ADAPTIVE TIME-FREQUENCY RESOLUTION IN VOCAL TRACT PARAMETER CODING FOR SPEECH ANALYSIS AND SYNTHESIS

A THESIS

Presented to

The Faculty of the Division of Graduate

Studies and Research

by

Charles Richard Patisaul

In Partial Fulfillment

of the Requirements for the Degree

Doctor of Philosophy

in the School of Electrical Engineering

Georgia Institute of Technology

June, 1974

ADAPTIVE TIME-FREQUENCY RESOLUTION IN VOCAL TRACT PARAMETER CODING FOR SPEECH ANALYSIS AND SYNTHESIS

Approved:

Aubrey M. Bush, Chairman

Thomas P. Barnwell, III

W. M. Leach

Date approved by Chairman: June 24, 1975

ACKNOWLEDGMENTS

Dr. A. M. Bush, my thesis advisor, introduced me to speech processing and provided me with invaluable guidance and encouragement during this research.

I owe much of my knowledge of speech and interactive computers to Dr. T. P. Barnwell, III. Without his day-to-day advice and assistance, this thesis would not have been possible. He also served on the proposal and reading committees.

Dr. J. C. Hammett, Jr. introduced me to the cepstrum vocoder and guided me through the early days of my involvement in speech processing.

Dr. W. M. Leach was a member of both the proposal committee and the reading committee. His comments and suggestions during the preparation of this dissertation were very helpful.

My wife, Nellie, cheerfully shared the ups and downs of graduate student life, all the while providing me with both moral and financial support.

This research effort was supported, in part, by the U. S. Army Research Office, Durham, North Carolina.

DEDICATION

To Nellie and Charlie

TABLE OF CONTENTS

TABLE OF CONTENTS (Concluded)

LIST OF TABLES

LIST OF ILLUSTRATIONS

LIST OF ILLUSTRATIONS (Concluded)

SUMMARY

Speech has been modeled as the response of a linear, time-invariant vocal tract filter to a stationary excitation which may be periodic or noise-like in nature. Vocoders make use of this model to code speech for efficient transmission by means of a deconvolution of the speech signal. Short-time spectral analysis is generally employed in the deconvolution procedure. A loss of resolution or detail in both time and frequency results from the spectral analysis and deconvolution.

A program of research was carried out to study the perceptual effects of reduced time-frequency resolution in the specification of the vocal tract filter and to test the utility of an adaptive resolution scheme suggested by the present knowledge of speech perception. This program involved the design and computer simulation of a cepstrum vocoder that adapted time and frequency resolution according to the voiced-unvoiced nature of the input speech. Speech processed by this vocoder was subjectively evaluated by a team of listeners using the Category Judgment Method. The criterion for evaluation was perceived quality.

Several tentative conclusions were drawn from the results of this research. It appears that time resolution on the order of 20.0 ms is adequate for vocoder applications. Systems that maintain this level of time resolution do not benefit from adapting to better time resolution in unvoiced regions and regions of transition between voiced and unvoiced speech. On the other hand, improvements in quality and/or reduc-

tions in data rates may be possible if systems that normally employ time resolution worse than about 20.0 ms make use of adaptive resolution.

Evidence that speech perception trades between time and frequency resolution was obtained by noting that reduction of frequency resolution in unvoiced and transition regions brings about no noticeable degradation in quality. The time-frequency trading notion was further supported by the observation that systems with the same data rate but different time-frequency resolution conditions may be judged equivalent in quality.

CHAPTER I

INTRODUCTION

The analysis and synthesis of speech signals have been areas of growing interest in recent years. Much of this interest has been directed toward data rate compression for speech transmission. Many speech compression techniques use analysis of the speech signal to extract vocal tract parameters which may be encoded more efficiently than the speech waveform. These parameters, along with parameters of the vocal tract excitation, are transmitted with a substantial savings in data rate. At the system receiver an approximation to the original speech is synthesized on the basis of the received parameters. Devices which perform this speech analysis and synthesis for compression are called vocoders (voice coders).

This dissertation reports research into one of the problems encountered in speech analysis-synthesis compression schemes, namely the loss of time and/or frequency resolution inherent in the analysis-synthesis procedure. The objective of this work was to study the effect of degraded time and frequency resolution which occurs in the specification of the vocal tract parameters on the perceived quality of processed (compressed and reconstructed) speech.

Chapter II presents a summary of the theory of speech production. The classical model for speech production is described. The concept of

speech deconvolution is introduced and four methods of speech deconvolution are reviewed. The last portion of Chapter II contains a discussion of the role of temporal and spectral features of the speech signal in the perception of speech and describes an experimental program to investigate the consequences of reduced time-frequency resolution in vocal tract coding in a perceptual environment. A strategy for adapting time-frequency resolution according to a voiced-unvoiced decision is also outlined.

In Chapter III the design and computer simulation of an adaptive cepstrum vocoder is described. The simulation employed a variety of time-frequency resolution conditions in the calculation of a vocal tract coding. The subjective evaluation of the resulting processed speech with respect to quality is discussed.

The results of a particular quality evaluation of speech which was processed by the adaptive cepstrum vocoder are interpreted in Chapter IV. Several tentative conclusions concerning both fixed and adaptive time-frequency resolution of vocal tract specification as they relate to speech quality are advanced.

Chapter V outlines recommendations for further work in several areas. These areas include extensions of the present work and consideration of new problems encountered in this research.

CHAPTER II

BACKGROUND

Analysis-synthesis data rate reduction techniques have been developed in order to obtain codings for speech information that are more efficient than direct representation of the speech waveform [1]. These techniques incorporate constraints which can be derived from the speech production mechanism and, to some extent, from the speech perception process. In this way advantage can be taken of the fact that the signal of interest is speech.

This chapter presents a review of the theory of speech production and describes the classical model for the speech mechanism. Four speech analysis-synthesis schemes based on this model are discussed and the problem of reduced time-frequency resolution is established. The process of speech perception is examined with attention to the role of temporal and spectral detail in perception. Finally, a program to further investigate their role is proposed.

Speech Production [1]

Speech results from the flow of air from the lungs through the glottis (the opening between the vocal cords), the throat, the mouth, and the nasal cavity. This flow of air serves as an acoustic excitation for the vocal tract. The vocal tract is an acoustic resonant cavity whose characteristics may be altered by changing the positions of the movable

parts of the tract, e.g., the tongue, soft palate (velum), jaw, and lips. These moving parts are called articulators. The resonances of the vocal tract are called formants.

Voiced speech results from periodic excitation of the vocal tract. For voiced speech, the glottis vibrates (i.e., opens and closes) producing a quasi-periodic flow of air that is rich in harmonics.

Unvoiced speech results from the excitation of the vocal tract by turbulent flow past a point of constriction in the tract or by the abrupt release of pressure at a point of constriction resulting in an initial burst followed by turbulent flow. Unvoiced excitation may be regarded as acoustic noise. Voiced and unvoiced excitation may occur simultaneously.

The sounds of General American (GA) English can be classified in six categories, namely

1. vowels
2. fricative consonants
3. stop consonants
4. nasal consonants
5. semi-vowels and glides
6. dipthongs and affricates.

These sounds may be represented in terms of phonemes which are the smallest distinguishable units of speech. The different manifestations that are identified as a particular phoneme are called allophones of that phoneme.

Vowels

Vowel sounds are produced by voiced excitation of a relatively stable vocal tract configuration. There is little constriction of the oral cavity. Radiation of sound is from the mouth. Since vowel production does not require motion of the articulators, they may be uttered as sustained sounds and are called continuants.

/i/ eve	/ɝ/ bird	/u/ boot
/I/ it	/ə'/ over (unstressed)	/U/ foot
/e/ hate*	/ʌ/ up	/o/ obey*
/ɛ/ met	/ə/ ado (unstressed)	/ɔ/ all
/æ/ at		/a/ father

(*These sounds usually appear as dipthongs in GA English.)

Fricatives

Fricative consonants result from unvoiced excitation of the vocal tract. Voiced excitation may occur in conjunction with the unvoiced excitation, producing a voiced fricative. Radiation of a fricative is normally from the mouth. Fricatives are also continuants.

Voiced	Unvoiced
/v/ vote	/f/ for
/ð/ then	/θ/ thin
/z/ zoo	/s/ see
/ʒ/ azure	/ʃ/ she
	/h/ he

Stop Consonants

Stop consonants are characterized by an abrupt release of pressure

at a point of constriction followed by turbulent air flow. Stop consonants thus require rapid motion of the articulators and are not continuant sounds. A stop may also be produced with simultaneous voicing.

Voiced	Unvoiced
/b/ be	/p/ pay
/d/ day	/t/ to
/g/ go	/k/ key

Nasal Consonants

The nasal consonants are voiced sounds produced with the velum open. There is complete closure at some point in the oral cavity so that radiation occurs through the nostrils. Nasals are continuant sounds.

/m/ m

/n/ no

/ŋ/ sing

Semi-Vowels and Glides

Semi-vowels and glides are sounds which closely resemble vowels. Semi-vowels are continuant, voiced sounds which are characterized by greater constriction of the oral cavity than in the case of vowels. Glides are dynamic sounds which precede vowels and which show movement toward the vowel.

/j/ you

/w/ we

/r/ real

/l/ let

Dipthongs and Affricates

Dipthongs and affricates are combination sounds produced by vocal

tract motion between phoneme positions. Dipthongs are vowel-like sounds characterized by a change from one vowel position to another. Similarly, affricates represent motion between certain stop consonant and fricative positions.

/eI/	say	/aI/	I
/Iu/	new	/oU/	go
/ɔI/	boy	/tʃ/	chew
/aU/	out	/dʒ/	jar

The phonemic description of speech is a linguistic abstraction that attempts to classify speech sounds in terms of fundamental units which cannot be further subdivided. These units or phonemes are combined to produce morphemes, the smallest sounds that have meaning in the language. Morphemes are then arranged to make up words, phrases, etc.

The GA phonemes may be thought of as a code relating specific vocal tract configurations and excitations to specific speech sounds. Such a viewpoint would suggest that speech can be discretized. In reality, in connected speech there is nearly continuous motion of the articulators from one sound to another as well as changes in the nature of the excitation. In some cases a mere vocal gesture toward a configuration for a phoneme is sufficient to convey that phoneme.

The key point of this discussion is that some sounds of speech are produced by relatively stable vocal tract configurations while others require rapid motion of the articulators and are thus transitory in nature.

The Speech Production Model

The theory of speech production has led to a model for the speech

mechanism which has proved to be of great utility [2]. This model has been especially useful in the design of analysis-synthesis data compression systems or vocoders. This model is shown in Figure 1. Here the speech signal s(t) is regarded as the response of a linear vocal tract filter with impulse response h(t) to the excitation e(t). For voiced speech, e(t) is a train of glottal pulses. For unvoiced speech, e(t) is acoustic noise.

A simplification results if the excitation is restricted to trains of impulses. For voiced excitation, e(t) is a train of impulses with a period equal to the period of the voicing. For unvoiced excitation, e(t) is a train of uniformly spaced impulses with random polarity. This is not an unreasonable restriction since we may think of the vocal tract filter as containing a "pulse-shaping" section which converts the idealized pulses into the actual excitation wave-shape. This is illustrated in Figure 2.

The preceding choice of excitation is of particular significance in the synthesis of speech. This follows since s(t) is the convolution of e(t) with h(t). Convolution involving impulse trains is computationally simple. Also, an analyzer based on this model need be concerned with the exact nature of the excitation only insofar as the pulse-shaping section affects the total impulse response h(t).

It should be noted at this point that the model under consideration makes no allowance for simultaneous voiced and unvoiced excitation. Hence it cannot properly handle certain speech sounds, e.g., the voiced fricatives. Simple superposition of the two types of excitation at the input of the vocal tract filter is not correct since, in general, the excitations

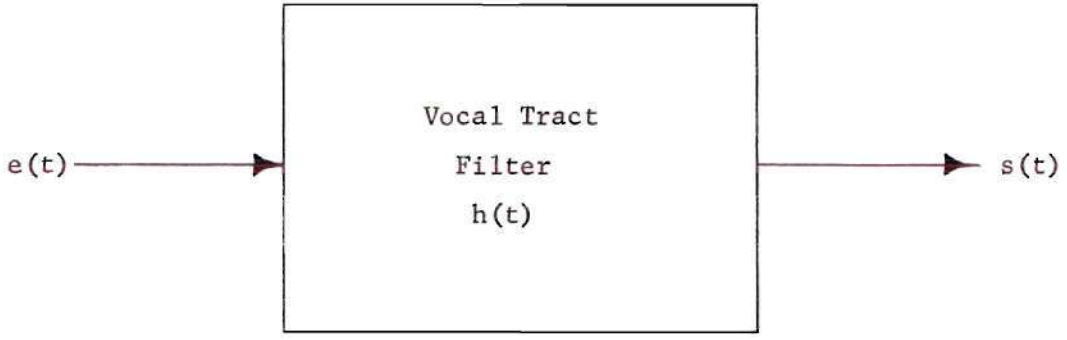


Figure 1. The Speech Production Model

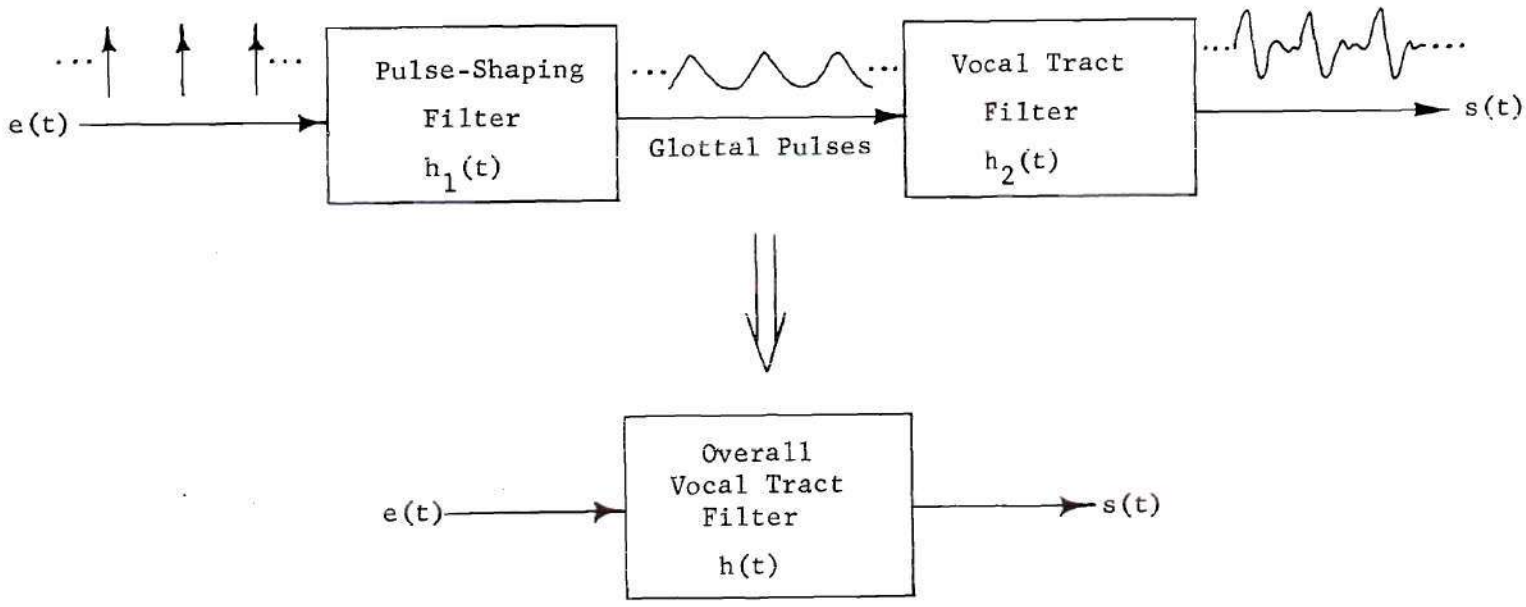


Figure 2. The Speech Production Model with Pulse-Shaping Filter

occur at different points in the tract and "see" different vocal tract filters. A more nearly correct approach is the superposition of the outputs of two separate vocal tract filters, one responding to voiced excitation, the other responding to noise excitation.

During the production of speech, both the configuration of the vocal tract and the nature of the excitation change with time. Thus the vocal tract filter and its excitation are time varying. Generally these variations occur relatively slowly (i.e., compared to the period of voiced excitation) with the important exceptions of stop consonant production and opening and closing of the velum to introduce and remove nasalization.

In order to simplify the analysis of the speech signal it will be assumed that $s(t)$ is the concatenation of segments produced by fixed excitations and stable vocal tract configurations. This assumption is valid if the segments are sufficiently short. In other words, $e(t)$ and $h(t)$ are assumed to be stationary in the short-term sense (i.e., their descriptions do not change over short periods of time). With this assumption the analysis of the speech waveform reduces to the successive analysis of time-invariant linear systems responding to stationary excitations.

Consider the stationary speech model with voiced excitation as shown in Figure 3. Because $e(t)$ is a periodic train of impulses in time with period T_o, the amplitude spectrum of $e(t)$ is periodic in frequency with period $1/T_o$. Thus the amplitude spectrum of $s(t)$ is a periodic line spectrum with an envelope determined by $|H(f)|$, the amplitude spectrum of $h(t)$.

In the case of unvoiced excitation $|E(f)|$ is not a line spectrum and hence $|S(f)|$ is not a line spectrum.

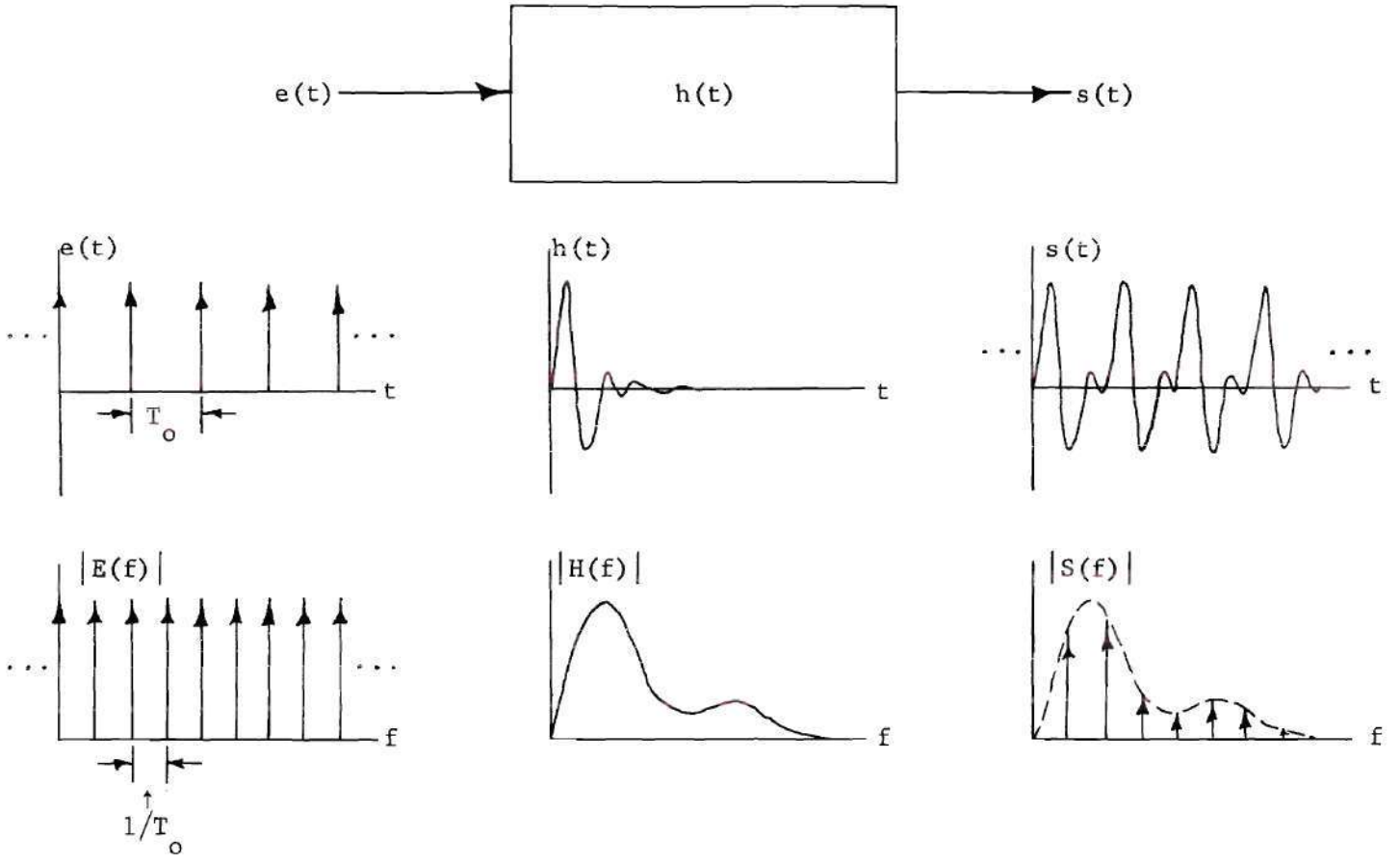


Figure 3. Idealized Waveforms for Voiced Speech

Deconvolution and Short-time Spectral Analysis of Speech

The central strategy of almost all vocoder systems is the decomposition of the speech signal s(t) into its component parts e(t) and h(t), or equivalently, E(f) and H(f). Hence the speech signal is deconvolved. With this deconvolution it is possible to extract pertinent parameters of e(t) and h(t). Since they are assumed to be slowly varying in nature, these parameters can be transmitted at a considerable savings of data rate over the rate required to transmit the original speech waveform. At the vocoder receiver, an approximation to s(t) is synthesized on the basis of the received excitation and vocal tract parameters. This approach was pioneered by Dudley with his invention of the channel vocoder in 1939 [3]. This concept is shown in block diagram form in Figure 4.

The analysis of speech based on the preceding speech model requires the analysis of short segments for which the stationarity assumption is valid. To this end, s(t) is multiplied by a so-called window function w(t) which is non-zero only over a period of time for which s(t) can be considered stationary. Let the windowed signal be defined

$$s_\tau(t) = s(t) \cdot w(t - \tau) \tag{2-1}$$

where the parameter τ determines the portion of s(t) which is windowed. The effect of this windowing in the frequency domain is a smearing of the spectrum according to

$$S_\tau(f) = S(f) * W(f,\tau) \tag{2-2}$$

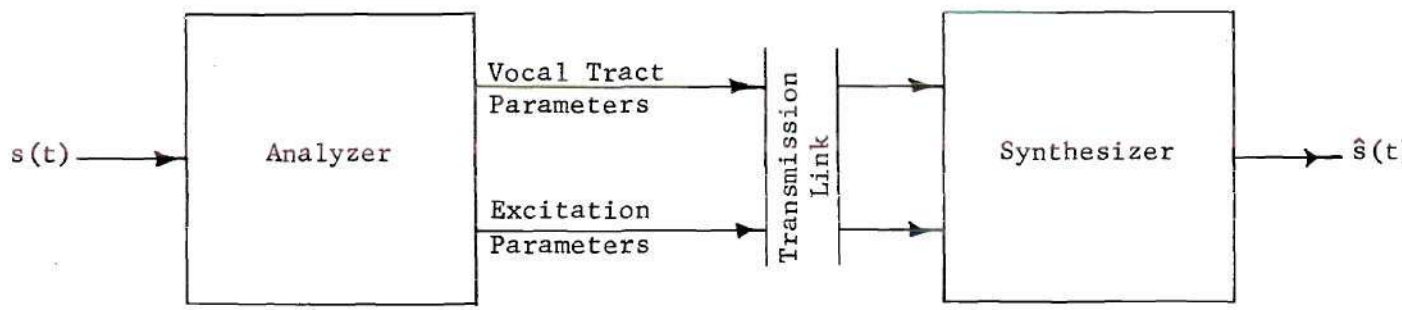


Figure 4. Conceptual Diagram of a Vocoder

where * denotes convolution and $W(f,\tau)$ is the Fourier transform or spectrum of $w(t - \tau)$. $S_\tau(f)$ is called the short-time spectrum of $s(t)$ in the vicinity of $t = \tau$ (assuming that $w(t)$ is non-zero around $t = 0$). Band-limited short-time spectra can be efficiently computed digitally by means of the fast Fourier transform (FFT) algorithm [4].

Most vocoders employ some form of short-time spectral analysis to approximate $e_\tau(t)$ and $h_\tau(t)$ or equivalently $E_\tau(f)$ and $H_\tau(f)$ where the subscript τ denotes stationary quantities associated with the speech segment $s_\tau(t)$. This type of approach is supported by evidence that, at an early stage of processing, the ear makes a crude spectral analysis [1]. Furthermore, Flanagan states:

> The peripheral auditory analysis made by the human cochlea is such that features of the short-time spectrum of the input signal are preserved. This analysis preserves temporal detail relevant to changes in spectral distribution, periodicity (or non-periodicity) and intensity. [1]

In the following four sections, four different vocoder strategies will be outlined. Each of these vocoders makes use of a form of short-time spectral analysis to derive an efficient coding for vocal tract parameters. The coding of the vocal tract may be derived in a variety of ways. Typically the excitation is coded in terms of a binary voiced-unvoiced decision and a measurement of the fundamental frequency or period of voiced excitation. The extraction of the excitation parameters is a problem in its own right and will not be considered in detail in this dissertation.

In order to establish a measure of the compression obtained by these techniques, it will be assumed that the speech signal occupies the frequency range from 300 to 3000 Hz. This is approximately telephone

bandwidth. For digital transmission of this signal, the sampling theorem [5] requires a sampling rate of 6000 samples per second. If these samples are quantized to eight bits (i.e., 256 levels) the resulting data rate is 48,000 bits per second (bps).

The Channel Vocoder

The notion of analysis-synthesis speech compression was introduced by Dudley [3] with his invention of the channel vocoder in 1939. A simplified diagram of Dudley's vocoder is shown in Figure 5. A short-time spectrum analysis is performed by the bank of bandpass filters, rectifiers, and lowpass filters spaced across the speech band. The output of each lowpass filter is the slowly varying envelope of the corresponding bandpass filter output. The lowpass filter outputs taken together form an approximation to the envelope of $|S_\tau(f)|$ or equivalently an approximation to $|H_\tau(f)|$. A voiced-unvoiced decision is made and a fundamental frequency value extracted in the voiced case. The excitation parameters are lowpass filtered to retain only the slowly varying features.

At the receiver or synthesizer, the appropriate excitation is generated and fed to a bank of bandpass filters identical to those in the analyzer. The level of the excitation applied to a given filter is controlled by the envelope signal received from the corresponding analyzer filter. The outputs of the bandpass filters are combined and an approximation $\hat{s}(t)$ of the original signal is produced.

The outputs of the 10 lowpass filters can be transmitted in a bandwidth of 250 Hz so that the bandwidth compression achieved is approximately ten to one. Tierney, et al. [6] have reported an analog channel vocoder

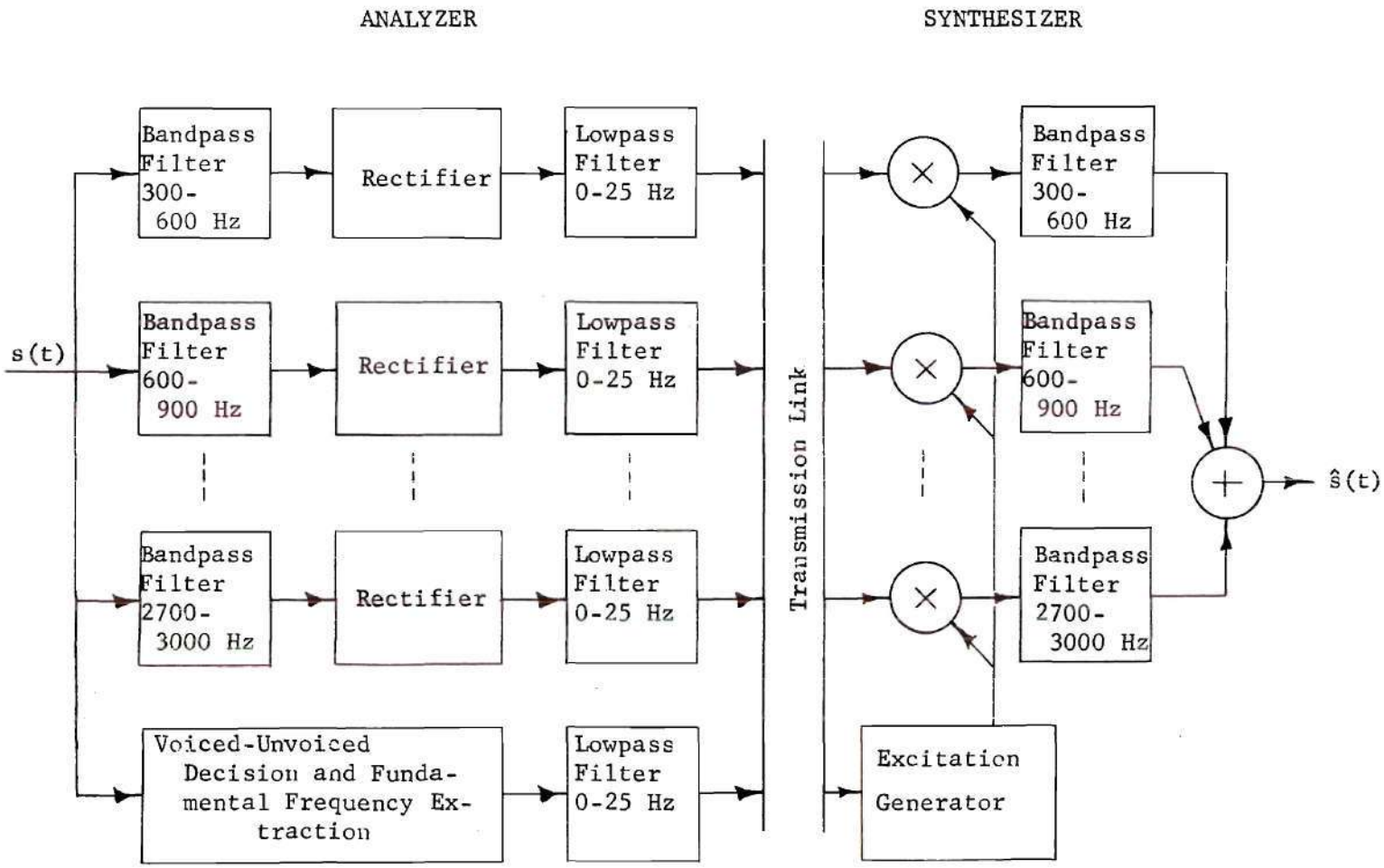


Figure 5. The Channel Vocoder

employing 18 bandpass filters covering the frequency range from 180 to 5100 Hz. The channel and excitation parameters were digitally multiplexed with a resulting transmission data rate of 5200 bps.

The synthesizer of the channel vocoder may be excited by a lowpass-filtered version of the original speech [1]. This filtered signal is "spectrally flattened" before it is presented to the bandpass filters. Vocoders excited in this manner are called voice-excited vocoders. This technique eliminates the need for an excitation extractor but requires greater bandwidth for transmission.

The channel vocoder may be implemented directly in digital form by the use of digital filtering techniques. A second approach to digital realization is the computation of the short-time amplitude spectrum by means of the FFT. The bandpass filtering may then be simulated by sampling bands of the computed spectrum.

The Formant Vocoder

The philosophy of the formant vocoder is to code the parameters of the vocal tract in terms of the formants or spectral peaks of $|H_T(f)|$. Figure 6 shows one type of formant vocoder. A short-time spectral analysis is performed and the frequencies and amplitudes of three formants (corresponding roughly to a frequency range of 0 to 3000 Hz) are determined. These frequencies and amplitudes, along with excitation parameters, are transmitted.

The synthesizer consists of three variable resonators controlled by the received formant frequencies and amplitudes. The excitation generator produces an excitation based on the received excitation parameters.

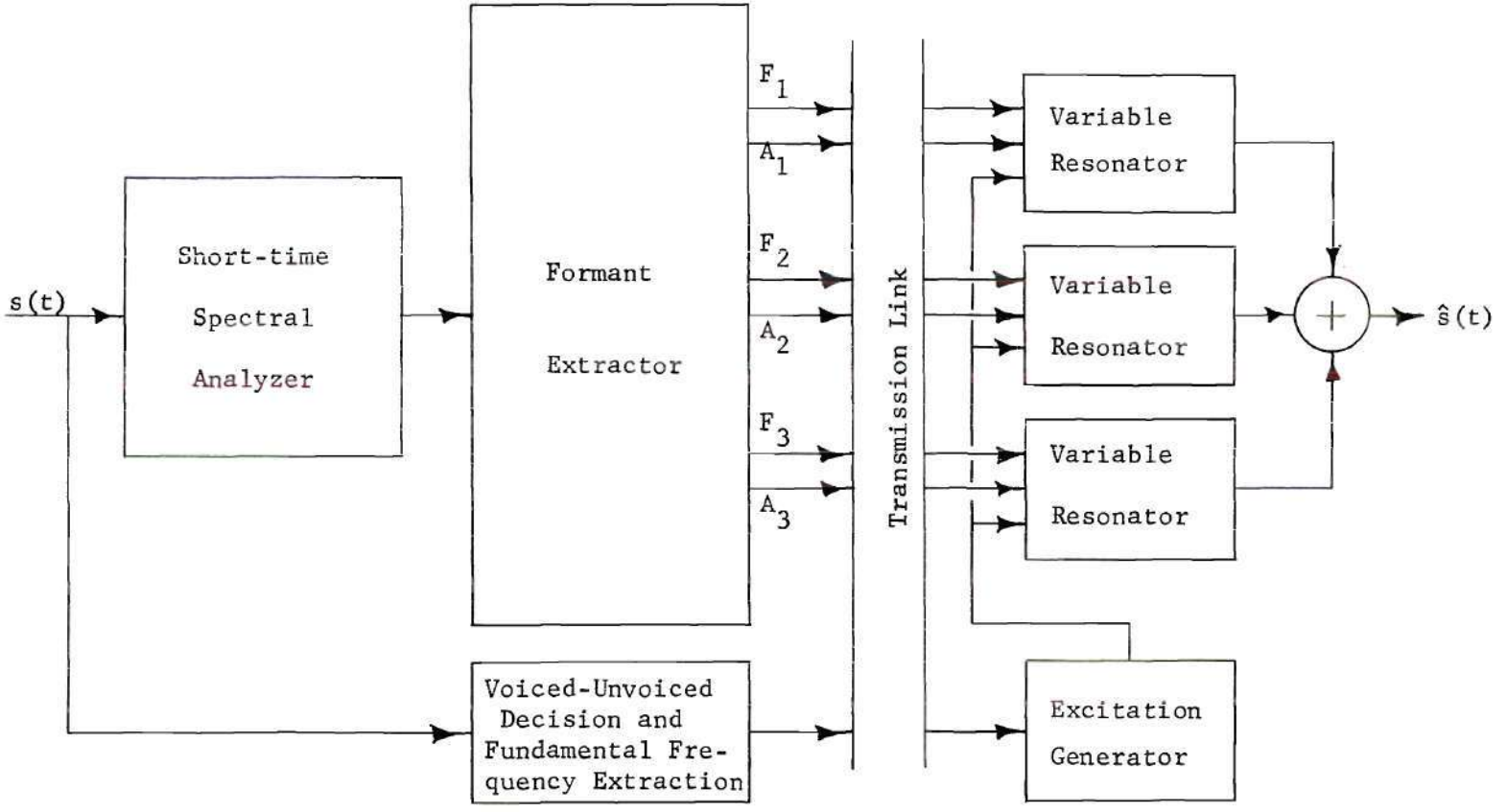


Figure 6. The Formant Vocoder

The resonances are driven with this excitation and their outputs combined to form the reconstructed signal $\hat{s}(t)$.

Another type of formant vocoder realization makes use of a cascade connection of the variable resonances. Both types may make use of more than three formants.

The formant vocoder approach works well for purely voiced speech. However, for unvoiced speech, the concept of formants seems to break down and formant vocoder performance is degraded.

Formant vocoders have great potential for speech compression (bandwidths of several hundred Hertz or data rates of 1000 bps or less). Both analog and digital versions have been implemented. The chief difficulty in formant vocoder realization is reliable and accurate automatic tracking of the formants [1].

The Linear Predictive Vocoder

The linear predictive vocoder [7] is an inherently digital vocoder based on the all-pole digital filter model given by

$$s(nT) = \sum_{k=1}^{K} a_k \, s((n - k)T) + e(nT) \qquad (2\text{-}3)$$

where $s(nT)$ is the n^{th} sample of the speech signal $s(t)$ sampled at a rate of $1/T$ samples per second. $e(nT)$ is the sampled version of the excitation $e(t)$. The a_k's are the filter coefficients or weights. This model is shown in Figure 7. Note that this model has no zeros. However, the effects of zeros over the frequency range of interest can be approximated by the all-pole model.

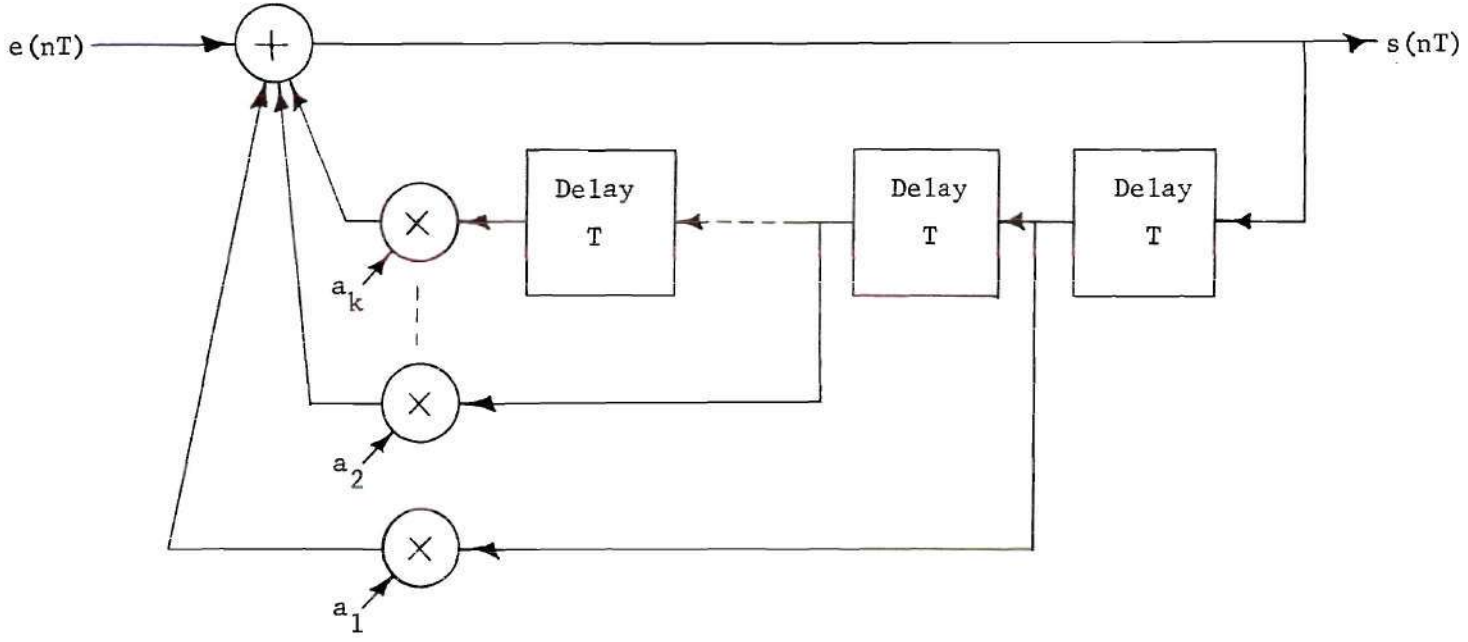


Figure 7. The Linear Predictive Speech Model

The strategy of the linear predictive vocoder is to estimate the a_k's at the analyzer and to transmit them, together with excitation parameters and gain factor G. The synthesizer consists of a digital filter controlled by the a_k's. This filter is excited by a reconstructed excitation sequence producing $\hat{s}(nT)$ at its output. The diagram of this vocoder is shown in Figure 8.

The analyzer estimates the a_k's by predicting the current speech sample on the basis of K previous samples according to

$$\tilde{s}(nT) = \sum_{k=1}^{K} a_k \, s((n - k)T) \tag{2-4}$$

where $\tilde{s}(nT)$ is the predicted value of $s(nT)$. This prediction explicitly ignores the excitation term in Equation (2-3). The a_k's are chosen so that the mean square prediction error $\overline{[s(nT) - \tilde{s}(nT)]^2}$ is minimized over some analysis interval $n_1 \leqq n \leqq n_2$. The a_k's are the solutions to the set of simultaneous equations

$$\sum_{k=1}^{K} \varphi_{jk} \, a_k = \varphi_{jo} \, , \qquad j = 1, 2, \ldots, K \tag{2-5}$$

where

$$\varphi_{jk} = \sum_{n=n_1}^{n_2} s_{n-j} \, s_{n-k} \tag{2-6}$$

The gain factor is computed from the RMS level of the input speech sequence.

Note that Equation (2-6) does not use a windowed version of $s(nT)$. One consequence of using "unwindowed" speech is that the computed a_k's may yield an unstable digital filter. Stability can be guaranteed by using

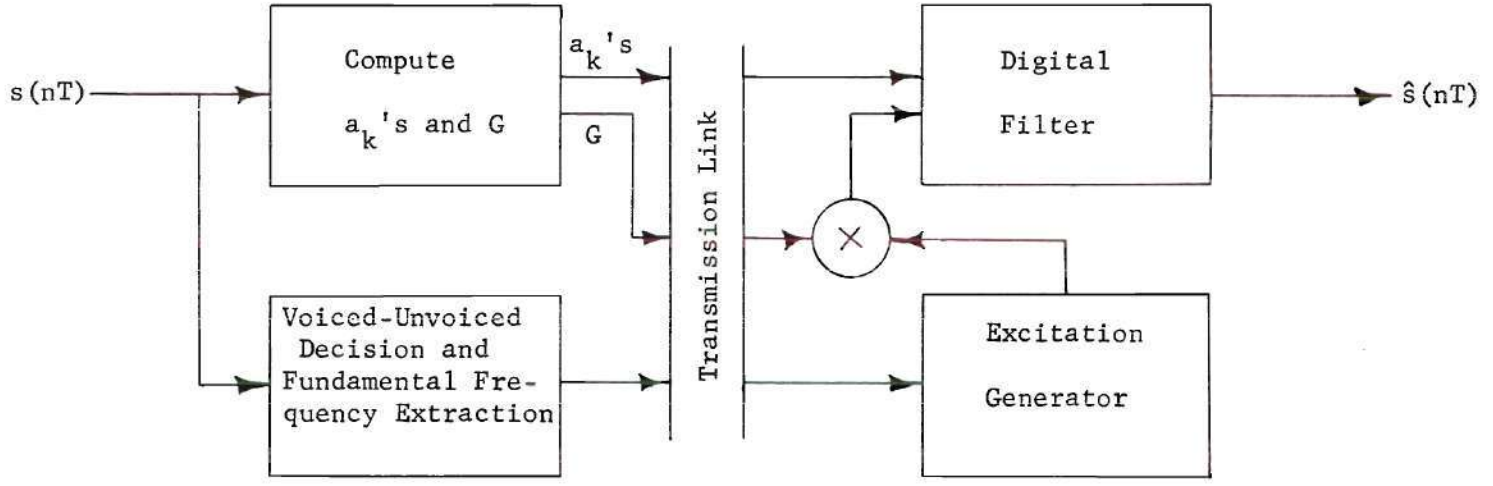


Figure 8. The Linear Predictive Vocoder

windowed speech in Equation (2-6) [46].

Typically eight to fourteen coefficients are transmitted for each 100 to 200 speech samples. Thus the data rate compression afforded by the linear predictive vocoder is obvious. Atal and Hanauer report good quality speech at bit rates around 4800 bps [7]. Markel has recently obtained good results at 3200 bps [8].

Although the linear predictive vocoder employs a time-domain analysis, it may also be thought of as performing a type of short-time spectral analysis since an approximation to the amplitude spectrum of the vocal tract filter can be calculated directly from the digital filter weights [4].

Several linear predictive vocoder realizations have been developed. The principal differences among them relate to the method for computing the a_k's.

The Cepstrum Vocoder

The cepstrum or homomorphic vocoder is based on a technique of speech deconvolution originated by Oppenheim [9,10]. The key to the operation of the vocoder is the transformation of the speech signal to a domain in which the effects of the excitation and vocal tract impulse response are additive. These can then be separated by linear filtering to accomplish deconvolution. A block diagram of the cepstrum vocoder is given in Figure 9. Typical cepstrum vocoder waveforms are shown in Figures 10 through 13.

With reference to Figures 9 through 13, the operation of the cepstrum vocoder is summarized in the following. The input speech is

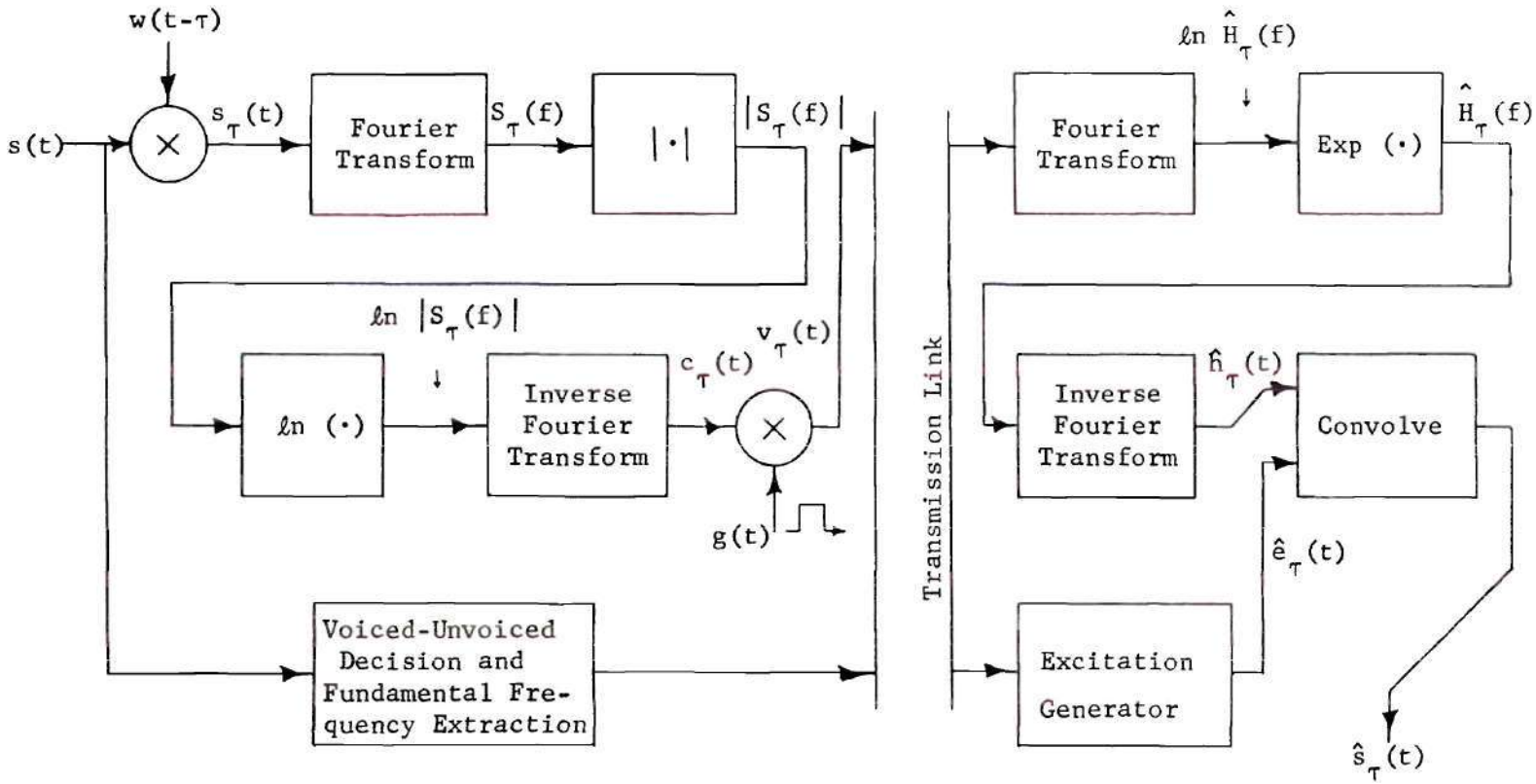


Figure 9. The Cepstrum Vocoder

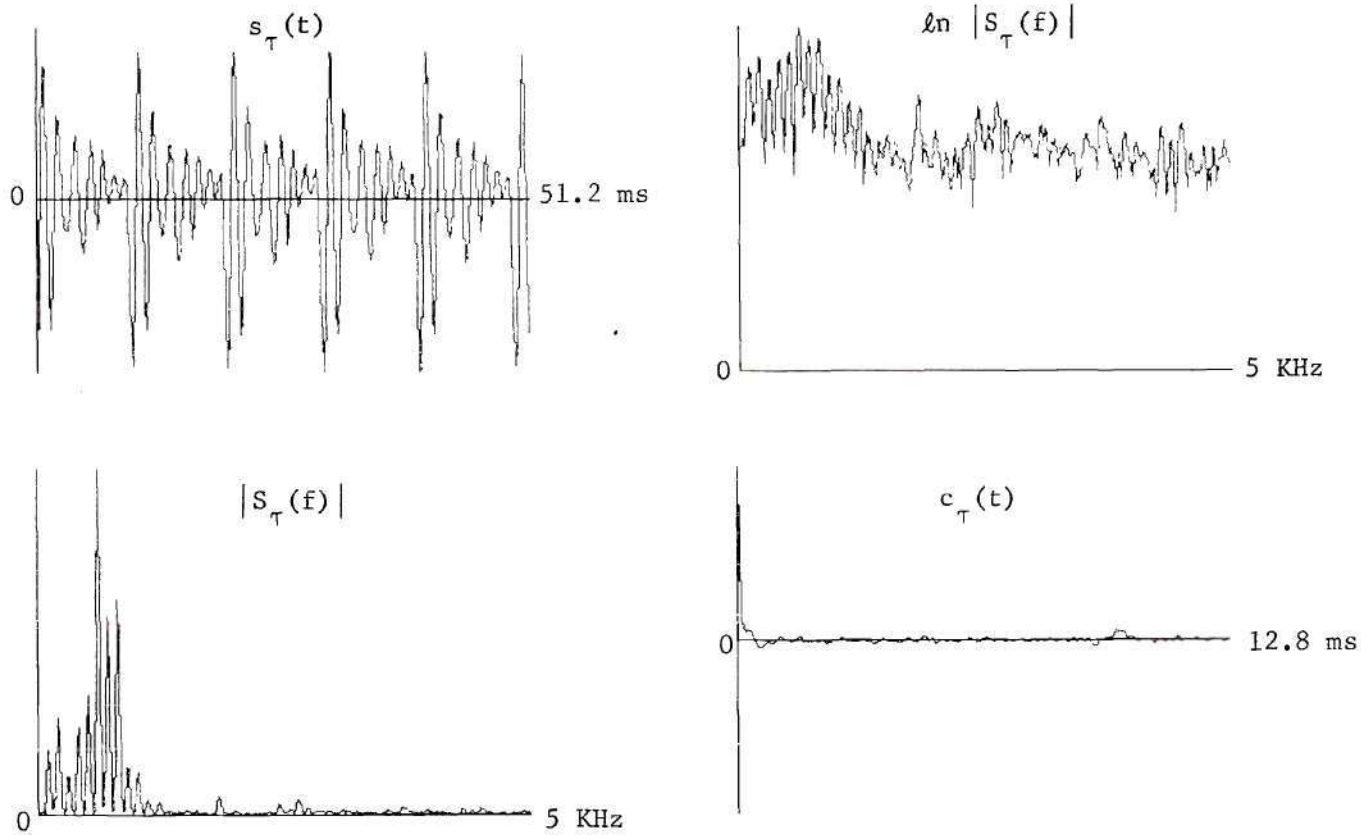


Figure 10. Cepstrum Analyzer Waveforms for a Sustained /a/

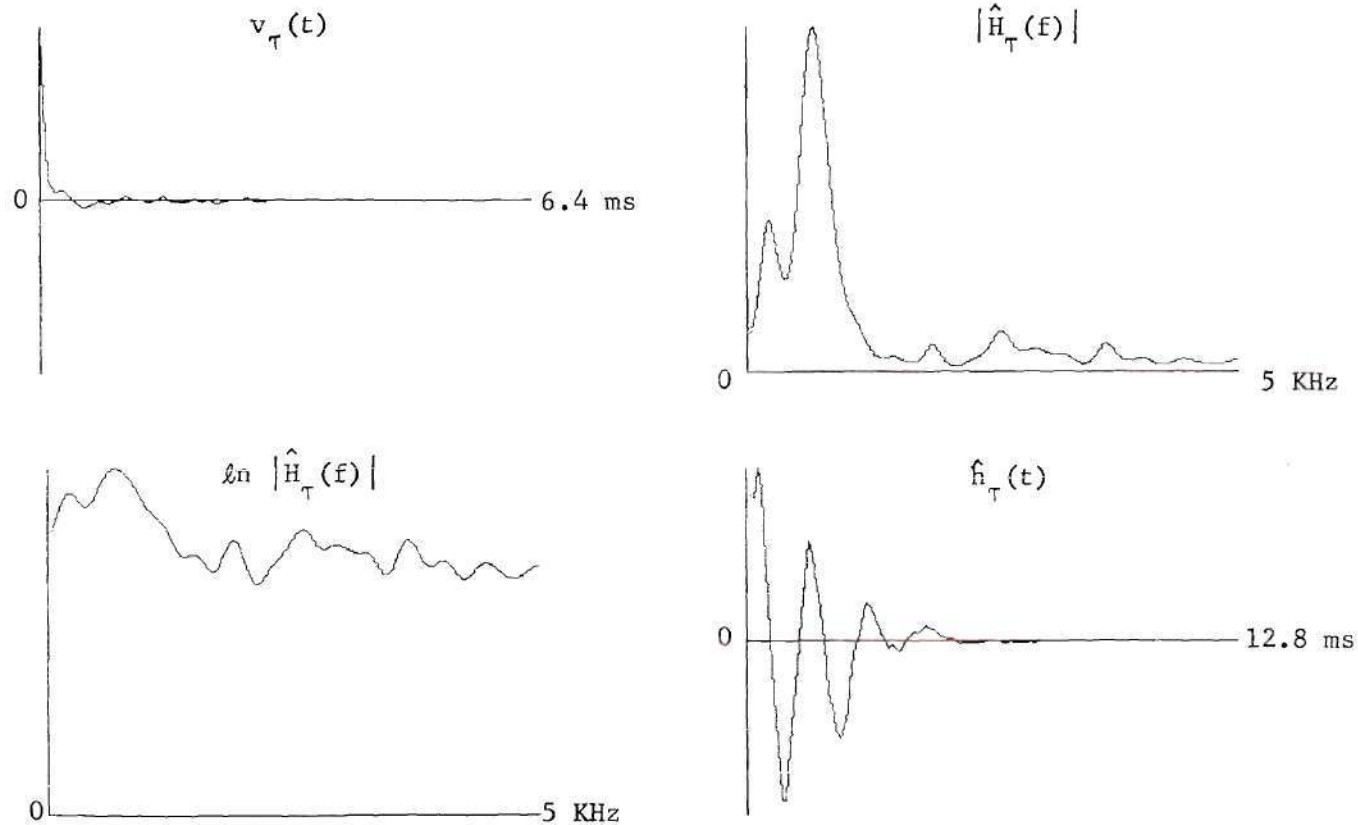


Figure 11. Cepstrum Synthesizer Waveforms for a Sustained /a/

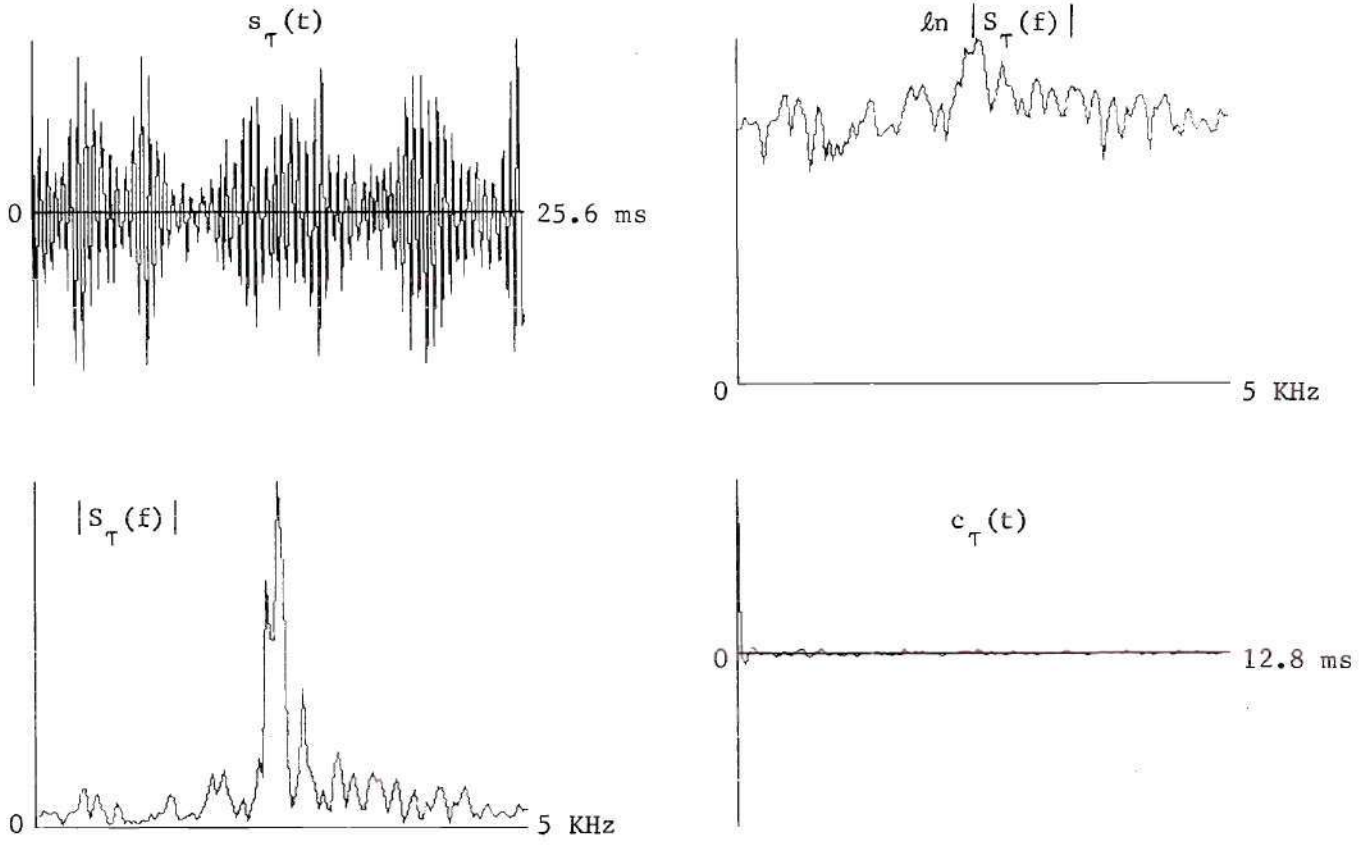


Figure 12. Cepstrum Analyzer Waveforms for a Sustained /ʃ/

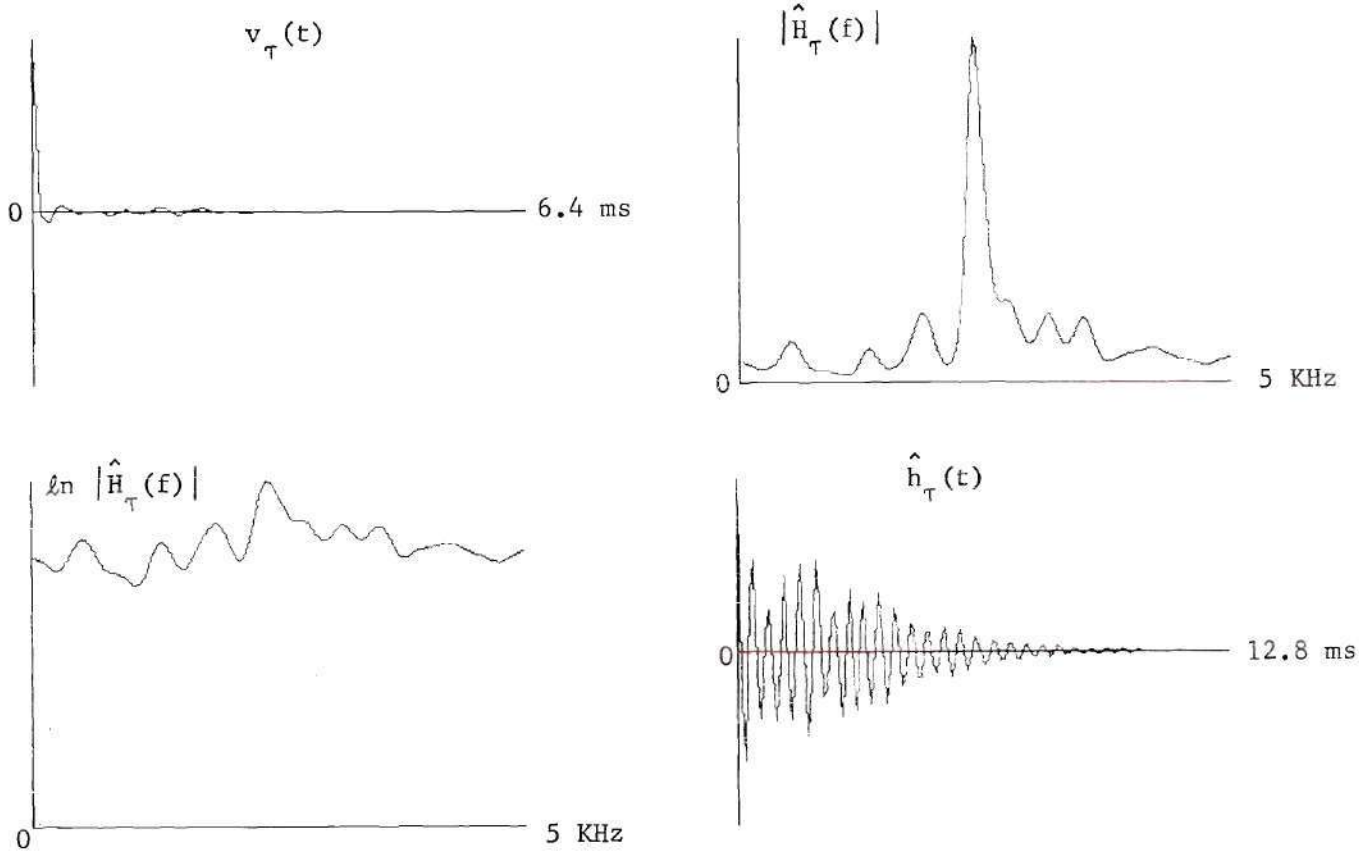


Figure 13. Cepstrum Synthesizer Waveforms for a Sustained /ʃ/

windowed by $w(t-\tau)$ so that the speech segment $s_\tau(t)$ may be considered stationary. The short-time spectrum $S_\tau(f)$ is then computed. $S_\tau(f)$ is approximately the product of $E_\tau(f)$ and $H_\tau(f)$. This approximation results from the influence of the window function. The specific effect of the window function will be ignored here and examined in detail in Chapter III. The magnitude of $S_\tau(f)$ is taken to yield

$$|S_\tau(f)| \cong |E_\tau(f)| \cdot |H_\tau(f)| \tag{2-7}$$

The natural logarithm is then computed, with the result

$$\ell n\ |S_\tau(f)| \cong \ell n\ |E_\tau(f)| + \ell n\ |H_\tau(f)| \tag{2-8}$$

At this point, the effects of the excitation and vocal tract response are additive and, under certain conditions, may be separated by linear filtering. To this end, the inverse short-time transform of $\ell n\ |S_\tau(f)|$ is computed giving the so-called cepstrum $c_\tau(t)$. For voiced speech $|H_\tau(f)|$ is a "smooth" function of frequency when compared to $\ell n\ |E_\tau(f)|$. Consequently the effect of the vocal tract response is largely confined to the low-time region of $c_\tau(t)$ while the effect of the excitation is largely confined to the high-time region of $c_\tau(t)$. If $c_\tau(t)$ is gated by $g(t)$ so that only the low-time features are retained, we have a coding of the vocal tract impulse response given by $v_\tau(t)$. If $s_\tau(t)$ is a voiced segment, then $|E_\tau(f)|$ is periodic in frequency with period $1/T_o$ where T_o is the period of the voicing. In this case, $\ell n\ |E_\tau(f)|$ is also periodic in frequency and a prominant spike occurs in the high-time region of $c_\tau(t)$. This spike

is located at $t = T_o$. For unvoiced excitation $|E_\tau(f)|$ is not periodic and no spike occurs in the high-time region of $c_\tau(t)$. In the case of unvoiced speech, truncation of the cepstrum yields a coding for $h_\tau(t)$ if the spectrum $|E_\tau(f)|$ is assumed to be flat over the frequency range of interest. Thus the cepstrum yields a coding for $h_\tau(t)$ and a source of information regarding the nature of the excitation. Noll [11] demonstrated a cepstral pitch and voicing detector based on detecting and measuring the location of the cepstral spike.

At the synthesizer $v_\tau(t)$ is Fourier transformed to give $\ell n\, \hat{H}_\tau(f)$. Exponentiation is performed to remove the logarithm of the analysis and the result is $\hat{H}_\tau(f)$, an approximation to the short-time spectrum of the vocal tract impulse response for the segment of speech $s_\tau(t)$. The inverse transform is computed to yield $\hat{h}_\tau(t)$, the synthetic vocal tract impulse response. The final step is to convolve $\hat{h}_\tau(t)$ with $\hat{e}_\tau(t)$ to produce $\hat{s}_\tau(t)$. Successive stationary segments of speech are processed by propagating the window along the speech waveform.

Note that this scheme discards all phase information at the analyzer. It has been shown that synthesis using a minimum-phase $\hat{h}_\tau(t)$ leads to synthetic speech of high quality [10]. Minimum-phase synthesis can be achieved by a simple weighting of $v_\tau(t)$. For a digital implementation, this weighting corresponds to forming the truncated cepstrum samples according to

$$v_\tau(nT) = \begin{cases} 2c_\tau(nT), & 1 \leq n \leq N \\ c_\tau(nT), & n = 0 \\ 0, & \text{otherwise} \end{cases} \qquad (2\text{-}9)$$

where $v_\tau(nT)$ is the n^{th} sample of $v_\tau(t)$, $c_\tau(nT)$ is the n^{th} sample of $c_\tau(t)$, T is the sample spacing in time, and NT is the length of the cepstrum truncation. Forming the truncated cepstrum in this manner causes the computed $\hat{h}_\tau(t)$ to be causal and a minimum-phase function; i.e., the phase of $\hat{H}_\tau(f)$ can be determined from the log magnitude through the Hilbert Transform [9,10].

In a digital implementation of the cepstrum vocoder, compression arises from the fact that only a few cepstrum samples are transmitted for many samples of input speech (e.g., 25 cepstrum samples for every 200 speech samples). Oppenheim has achieved high quality vocoder speech at 7800 bps [7].

Deconvolution and Time-Frequency Resolution

Each of the speech compression techniques presented in the preceding four sections attempts to code the vocal tract and excitation in such a way that the short-time amplitude spectrum of the original signal is in some sense preserved in the processed speech signal. If the speech signal could be segmented into perfectly stationary pieces and if these pieces could be exactly deconvolved, then the short-time amplitude spectrum could be faithfully reproduced. Unfortunately, connected speech is never perfectly stationary (i.e., never really time-invariant), and exact deconvolution is never achieved. Thus the short-time spectrum is preserved in neither time nor frequency.

Time resolution may be defined in terms of the shortest time event that can be resolved by an analyzer. Similarly, frequency resolution may be thought of as a measure of spectral detail retained in the short-time

amplitude spectrum.

The time resolution of the channel vocoder is determined by the memory associated with the analyzer and synthesizer filters since this memory determines the shortest time event that can be resolved. Roughly speaking, the memory of each filter is inversely proportional to its respective bandwidth. Frequency resolution in the channel vocoder is a function of the bandwidths of the bandpass filter and the spacing of their center frequencies across the speech band. The arrangement of the bandpass filters establishes the accuracy with which the short-time amplitude spectrum is represented in frequency. If sharp, narrow filters are spaced in a contiguous manner across the speech bandwidth, good frequency resolution results. On the other hand, if the analyzer filters are broad and "sloppy," frequency resolution is poor.

For the formant vocoder, time resolution is a function of the time window employed in the short-time spectral analysis that is done prior to formant extraction. Resolution in time is influenced both by the effective duration of the analysis window and by the manner in which the window is propagated along the speech waveform.

The frequency resolution of the formant vocoder depends on the accuracy of the formant extraction process and, more basically, on how well the short-time spectrum of the vocal tract can be represented by a fixed number of resonators.

Time-frequency resolution for the linear predictive vocoder may be discussed in much the same terms as for the formant vocoder. Time resolution is determined by the duration and propagation of the analysis window. The accuracy of the calculation of the filter weights and the

validity of the all-pole vocal tract model are factors contributing to the frequency resolution of the linear predictive vocoder.

Time window duration and propagation also determine the time resolving ability of the cepstrum vocoder. Deconvolution is accomplished by gating the cepstrum which corresponds to a linear smoothing of $\ell n\ |S_{\tau}(f)|$. The extent of this smoothing determines the frequency resolution of the cepstrum vocoder. Hence, frequency resolution is dependent on the length of the cepstrum retained after truncation. The time and frequency resolution properties of the cepstrum vocoder will be explored in greater detail in Chapter III.

Good time resolution and good frequency resolution are often conflicting goals in practical vocoder design. For example, there may be explicit interaction of factors influencing time and frequency resolution. This is the case with the channel vocoder. An attempt to improve frequency resolution by decreasing the bandpass filter bandwidths would have the effect of increasing the memory of the filters with a resulting loss of time resolution.

Time and frequency resolution may be in conflict in another way. Consider a vocoder operating at a fixed data rate or bandwidth. If one resolution is to be improved, then the other must be degraded in order to maintain the fixed data rate or bandwidth. For instance, if the frequency resolution of a linear predictive vocoder is improved by transmitting more filter coefficients, then the coefficients must be transmitted less often (poorer time resolution) in order to keep the same data rate. If the analysis window of a cepstrum vocoder is shortened to improve time resolution, the fixed data rate requires a more severe truncation of the

cepstrum and hence reduced frequency resolution.

Vocoders produce processed speech that is imperfect in both time and frequency. Since practical considerations place good time resolution and good frequency resolution in opposition, vocoders employ compromise resolution conditions. The perceptual effect of a given time-frequency resolution compromise depends upon the use of the temporal and spectral detail in the speech perception process. The perception process is examined in the following section with special attention given to the role of time and frequency resolution.

Speech Perception

The perception of speech is not understood nearly as well as its production. At present no complete theory of speech perception exists. However, there has been much study and experimentation that sheds light on speech perception. For the most part this work can be divided into two categories:

1. measurement of auditory perception
2. recognition of acoustic signals within a linguistic framework.

The first category deals primarily with determining the resolving power of the hearing mechanism. The second examines the identification and classification of auditory patterns which have significance with respect to the speech communicative experience of the listener.

Classical measurements of auditory perception usually involve an examination of discrimination of a single characteristic of the stimulus. The measurements are generally based on differential discrimination or close comparison. It seems unlikely that the complex central processing

involved in speech perception is brought into play in such measurements. Some of these experiments which may relate to speech are outlined below.

Flanagan [1] reports difference limens (DL's) or just noticeable differences (JND's) for some of the characteristics of vowels. These DL's are:

1. formant frequency -- 3 to 5%
2. formant amplitude -- 1.5 to 3 db
3. formant bandwidth -- 20 to 40%
4. spectral "valleys" -- +13 to -∞ db
5. pitch -- 0.3 to 0.5%

Morton and Carpenter [12] measured DL's for formant frequency which were in agreement with Flanagan. They also determined that changes in the relative intensity of the two most prominent harmonics may be a perceptual cue to a shift in formant frequency.

Moore [13] performed experiments to determine the frequency DL's for short-duration tonal pulses. He discovered that the DL's varied with both tone duration and absolute frequency, suggesting a sort of temporal-spectral trading in the auditory perception process.

Ronken [14] has studied the effects of bandwidth and duration on frequency discrimination. He measured the discriminability of short-duration tones having equal bandwidth but different time envelope duration. The tones were not equally discriminable but rather their discriminability was time envelope dependent. The conclusion is that temporal aspects of the signal influence frequency discrimination.

In an experiment to measure the ability of listeners to detect a

gap between tonal pulses of differing frequency, Williams and Perrott [15] found that the smallest detectable gap was a function of pulse duration and frequency difference between the pulses. The gap resolution became worse with increasing pulse duration and increasing frequency differences.

In a study of broadband noise, Malme [16] observed that spectral peaks with Q's of less than about five in an otherwise flat spectrum were not discernible. Spectral notches with Q's of less than about eight were not perceptible.

Brady, House, and Stevens [17] carried out an experiment with a single time-varying formant. The variable formant was excited with a short burst of a pulse train while the formant frequency underwent a rapid (20 ms) transition. Listeners were asked to adjust a non-varying resonator so that it sounded as much like the test signal as possible. There was a strong tendency to adjust the frequency of the fixed resonator to the final frequency of the varying formant. This tendency was especially strong when the formant transition occurred early in the test signal.

Experiments such as the ones just described may not be particularly relevant to speech perception. However, their results should establish an upper bound on the importance of time and frequency resolution in the perception of speech since they attempt to measure the maximum resolving ability of the auditory mechanism.

Speech is a multi-dimensional signal whose perception requires the identification of the sounds of speech and an association of these sounds with linguistic elements. Evidence exists that some sounds are perceived as individual phonemes while others require more clues. For example, the nasal consonants are difficult to distinguish in isolation but are easily

identified in normal connected speech [1]. In this vein, a large body of experimental work has been done utilizing single syllables composed of a few phonemes produced either synthetically or naturally. The main objective of this work has been the determination of the acoustic features which serve as cues to phoneme recognition.

Pyron and Williamson [18] studied the distribution in time of stop consonant recognition cues. This study was accomplished by the gating out of certain portions of the time waveform. Their results indicate that recognition cues for certain stop consonants are concentrated heavily in the initial 10-15 ms of the acoustic signal.

An investigation of stop consonants by Halle, Hughes, and Radley [19] suggests that the presence of a burst, intensity of the burst, silence before or after the burst, and formant transitions before or after the burst are all strong cues in stop consonant perception. The presence or absence of voicing was not a strong cue. By studying bursts of stop consonants in isolation they were able to conclude that the stop can often be identified from its burst, an indication that the gross spectral shape of the stop is meaningful. By blanking the burst from syllables containing final stops, they observed that the consonants may also be identified by formant transitions in the preceding vowel. In conclusion, they postulate a time-frequency trading in speech perception with formant transitions as a middle ground.

Further support for the identification of stops from their burst is provided by the recent work of Winitz, Scheib, and Reeds [21]. They also propose that the vowel associated with the stop can sometimes be identified from the burst portion of the stop.

According to Liberman [20], the frequency range of frication noise gives clues to the identity of the stops and some of the fricatives. In addition, he states that formant transitions are important in distinguishing stops, nasals, and semi-vowels with the direction, extent, and duration of the transition being relevant features.

Sharf and Hemeyer [22] examined formant transitions relative to stop and fricative perception. By excising the noise portion of stops and fricatives from natural syllables, they were able to observe a strong dependence on formant transitions in the recognition of these consonants. They found a significant advantage of vowel-consonant transitions over consonant-vowel transitions. Also, voiced consonant identification from vowel formant transition was superior to that for unvoiced consonants.

Vowel formant transitions also play a role in the voiced-unvoiced distinction of stop consonants. Stevens and Klatt [23] report that the recognition of voiced and unvoiced stops in initial, prestressed positions is influenced by the voicing onset time after the release of the burst and by the presence or absence of rapid formant transitions at the onset of voicing. Short onset times coupled with rapid formant transitions lead to the identification of a voiced stop. On the other hand, long onset times and the absence of rapid transitions cause stops to be identified as unvoiced.

Formant transitions in the perception of semi-vowels have been studied by Ainsworth [24] who observed that alterations in the loci, shape, and duration of the transitions lead to the perception of different semi-vowels.

Preceding vowel duration has been found to influence the voiced-

unvoiced distinction of stops. Raphael [25] noted this effect in word-final stops, fricatives, and consonant clusters. He noted, in addition, that the vowel duration cue sometimes overrides other voiced-unvoiced cues that may be present. Denes [26] suggests that the voiced-unvoiced identification of a final fricative depends on the ratio of the duration of the fricative to the duration of the preceding vowels with perception tending toward voiced as the ratio decreases.

In a study of the recognition of synthetic vowels in isolation and in the "h __ d" context, Ainsworth [27] observed that listeners were influenced in their identification of the vowels by the duration of the vowel as well as by its formant frequencies. Another experiment indicated that the perceived duration of a vowel was affected by the durations of sounds immediately prior to the vowel.

Hughes and Halle [28] have experimented with segments of fricative consonants isolated from real speech. After presenting these segments to listeners for identification they concluded that some fricatives can be recognized in isolation, suggesting that the gross spectral shape of fricatives has perceptual significance.

Speech perception beyond the isolated syllabic level is poorly understood because of the increased complexity of the processing involved. The influences of vocabulary, grammatical structure, and context seem to become as important as the acoustic features of the speech waveform. The perceptual units of speech are not well defined and, in fact, seem to vary from single phonemes to syllables to words or even to whole phrases in some circumstances [1]. Several recent studies have pointed up the complexity of speech perception beyond the syllabic level.

In an investigation of the JND's for segment duration in natural speech, Huggins [29] altered the durations of certain segments imbedded in running speech. He then asked listeners to judge whether the segments were of normal, long, or short duration. The JND's were found to be phoneme and context dependent. In another study, Huggins [30] concluded that changes in the duration of some segments of natural connected speech required compensation in the duration of other segments in order to maintain temporally fluent speech. Compensation was not required in all cases but when it was, the need for compensation was greatest across word boundaries.

McNeill and Repp [31] estimated that the perception process for stop consonants in consonant-vowel syllables requires appreciably more time than the acoustic manifestation of the phoneme requires. They hypothesized that forward masking of the input to the perception process may occur from the onset of the stop consonant stimulus and that the masking may cross phoneme or syllable boundaries. The implication was that speech perception relies heavily on internal processing rather than on detailed attention to the acoustic signal and may, in fact, operate some of the time in the absence or attenuation of input. This hypothesis is supported by the fact that a variety of acoustic cues can lead to the perception of the same phoneme.

An earlier experiment by Ladefoged and Broadbent [32] suggests that vowel identification may be based not on absolute formant structure but rather on formant structure in relation to other vowels spoken by the same speaker.

Speech perception on the suprasegmental level depends heavily on the fundamental frequency of voicing. The shape of the fundamental frequency contour aids the listener in the determination of stress and intonation which allow the recognition of structure and semantic value in an utterance. Barnwell [47] discusses the role of fundamental frequency in the perception of stress and intonation. Fundamental frequency is a prosodic feature of the speech signal and hence cannot be studied in terms of vocal tract parameters.

Several conclusions can be drawn from this discussion of speech perception. First, speech perception is a complex and, at best, poorly understood process, particularly the perception of connected speech. Secondly, speech perception appears to make use of both temporal and spectral features of the speech signal and trades between the two to some extent. Last, temporal cues such as bursts, transitions, durations, and voicing onsets seem to play significant roles in the perception of many sounds, especially at the junctures of voiced and unvoiced speech. These conclusions form the motivation for the program of research reported in this dissertation.

The Research Program

The discussion presented in this chapter serves to motivate an experimental research program to examine the perceptual effects of imperfect time-frequency resolution in vocal tract specification in a vocoder context. Additional motivation comes from the work of Hammett [33].

The research involved the use of a vocoder to impose varying time and frequency resolution conditions on natural, connected speech. The use of running speech made this a study of the sort alluded to by

Huggins [30] when he wrote ". . . words produced in isolation are not equivalent to words produced in fluent speech, or vice versa, since they cannot be interchanged without a serious loss of intelligibility."

The apparent importance of temporal cues to the perception of speech in regions of transition between voiced and unvoiced excitation suggests that a vocoder might achieve higher quality speech at a given data rate if it employed improved time resolution in these areas. To test this hypothesis the vocoder used was designed to adapt its time-frequency resolution on the basis of a voiced-unvoiced decision.

The adaptive strategy is similar to approaches employed by Gold and Rader [34] and Hammett [33]. Gold and Rader designed a channel vocoder which adapted its time resolution on the basis of a spectral derivative, using better time resolution during periods of rapid spectral change. They were able to lower the bit rate of 2400 bps for the non-adaptive vocoder to about 1800 bps with the adaptive version without noticeable degradation in quality.

Hammett used a cepstrum vocoder which adapted its time-frequency resolution according to a cepstral derivative or distance measured in much the same manner as Gold and Rader. Using this strategy, he obtained good quality speech at 3700 bps.

The adaptive techniques of Gold and Rader and Hammett show potential for increased quality and decreased data rate but need more evaluation. The approach used here is somewhat less general than the spectral and cepstral derivative approaches but is more tractable and should serve well to further the study of adaptive resolution.

Consideration of frequency resolution has been included in this work because, as discussed earlier in this chapter, improvements in time resolution are often accompanied by reductions in frequency resolution.

The cepstrum vocoder was selected for this study for several reasons that will receive elaboration in Chapter III. The most important of these reasons was the ease and relative independence with which the time and frequency resolution of the cepstrum vocoder can be manipulated.

The cepstrum vocoder was computer-simulated using a variety of time-frequency resolution conditions in both the adaptive and nonadaptive modes with accurate excitation information supplied externally. The performance of the vocoder was judged by subjective evaluation of the processed speech by a team of listeners. The criterion for performance was the perceived quality of the speech.

The results of this work should serve as a guide in the design of future vocoders and should contribute to the understanding of the role of time and frequency in speech perception. The remainder of this dissertation is devoted to reporting the experimental phase of the research and the evaluation of the results.

Summary

Speech has been modeled as the response of a linear, time-invariant vocal tract filter to a stationary excitation which may be periodic or noise-like in nature. Vocoders make use of this model to code speech for efficient transmission by means of a deconvolution of the speech signal.

Short-time spectral analysis is generally employed in the deconvolution procedure. A loss of time and frequency resolution results from

the spectral analysis and deconvolution.

A program of research has been carried out to study the perceptual effects of reduced time-frequency resolution and to test the utility of an adaptive resolution scheme suggested by the present knowledge of speech perception. The remainder of this dissertation describes that research and its results.

CHAPTER III

DESIGN, SIMULATION AND EVALUATION OF THE ADAPTIVE CEPSTRUM VOCODER

The research that is the topic of this thesis was conducted to study the perceptual consequences of time and frequency resolution degradation in vocal tract specification in a vocoder environment and to evaluate further the potential of an adaptive resolution strategy. A cepstrum vocoder which adapts in time and frequency resolution according to a voiced-unvoiced decision was designed and simulated on a digital computer. The simulations employed a variety of resolution conditions in both the adaptive and nonadaptive modes. The performance of the vocoder was judged in subjective listening tests with quality as the criterion for judgment. The listening test data were analyzed and several conclusions were drawn regarding the perceived quality of the processed speech as a function of the time and frequency resolution parameters of the processor.

The cepstrum vocoder was chosen as the research vehicle for this work because of several advantages it offers. The principal two of these are:

1. The vocoder lends itself readily to digital simulation, and
2. Time and frequency resolution are easily and almost independently manipulated through variations in the window duration, window propagation (framing) interval, and cepstrum truncation.

This chapter begins with an analysis of the time and frequency resolution properties of the cepstrum vocoder. Following this analysis,

the details of the vocoder design and simulation are presented. The design and administration of the subjective listening tests and the data reduction and interpretation are discussed.

Time and Frequency Resolution Properties of the Cepstrum Vocoder

The time resolution of the cepstrum vocoder is determined by the duration of the window function and by the rate at which the window is propagated along the speech waveform. Typically the window is propagated in discrete steps or frames. If the frame interval is such that there is little or no overlap of the window into adjacent frames, the time resolution can be considered equal to the frame interval. This is because speech is produced at the synthesizer in segments equal to the frame interval and because a short event in time will affect only one frame of processed speech if the event appears in only one frame of input speech. The window durations and frame intervals used in this study meet this condition.

The window function chosen was the Hanning or raised-cosine window given by

$$w(t) = \begin{cases} 0.5\left(1 + \cos\frac{2\pi t}{D}\right), & |t| \leq D/2 \\ 0, & \text{otherwise} \end{cases} \qquad (3\text{-}1)$$

where D is the duration of the window. This is the same window used by Oppenheim and Schafer [9,10] and Hammett [33].

Deconvolution is accomplished by a linear smoothing of $\ell n|S_\tau(f)|$. This smoothing is achieved by low-time gating of the cepstrum with a gate function specified by

$$g(t) = \begin{cases} 1, & |t| \leq L \\ 0, & \text{otherwise} \end{cases} \tag{3-2}$$

where L is the truncation length. The smoothing obtained is described by the convolution of $\ell n|S_\tau(f)|$ with the spectrum of the gate function which is [5]

$$G(f) = \frac{L \sin 2\pi Lf}{\pi Lf} \tag{3-3}$$

Smoothing in the frequency domain also results from the initial windowing of the speech signal. Since

$$s_\tau(t) = s(t) \cdot w(t - \tau) \tag{3-4}$$

we have in the frequency domain

$$S_\tau(f) = S(f) * W(f,\tau) \tag{3-5}$$

For the Hanning window [5]

$$|W(f,\tau)| = \left| \frac{D \sin \pi Df}{2\pi fD\,[1 - (fD)^2]} \right| \tag{3-6}$$

Figure 14 shows the effect of the window function for three values of duration D. Note that the ideal line structure for voiced speech is smeared or smoothed by the spectrum of the window function, and that the smearing becomes more pronounced as the window is shortened.

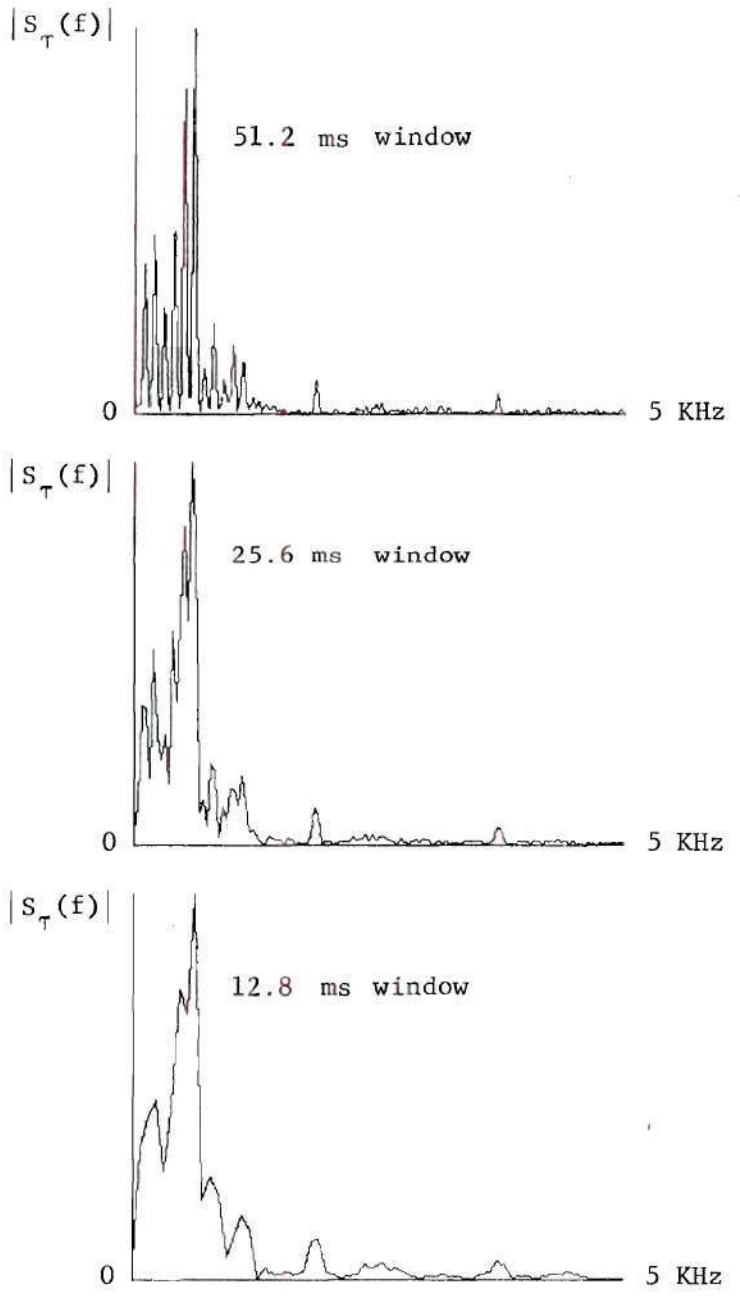


Figure 14. Frequency Domain Smearing Due to the Window Function for a Sustained /ʌ/

Equation (3-4) can be rewritten as

$$s_\tau(t) = \left[e_\tau(t) * h_\tau(t) \right] \cdot w(t - \tau) \tag{3-7}$$

Fourier transformation of this equation yields

$$S_\tau(f) = \left[E_\tau(f) \cdot H_\tau(f) \right] * W(f,\tau) \tag{3-8}$$

and

$$\ell n\left| S_\tau(f) \right| = \ell n\left| \left[E_\tau(f) \cdot H_\tau(f) \right] * W(f,\tau) \right| \tag{3-9}$$

The potential for deconvolution is not yet obvious from Equation (3-9).

The speech model adopted in Chapter II restricts the vocal excitation function to trains of impulses. Since this is the case the approximation

$$s_\tau(t) \cong \left[e_\tau(t) \cdot w(t - \tau) \right] * h_\tau(t) \tag{3-10}$$

may be made if w(t) is assumed to be smooth, i.e., essentially constant over the duration of $h_\tau(t)$ [9,10,33]. This approximation is equivalent to applying the window function to the input of the vocal tract filter rather than the output. This assumption may not be reasonable in many situations. However, there is evidence (to be pointed out later) that good results can be obtained even when the assumption is apparently violated.

With this assumption, $\ell n|S_\tau(f)|$ can now be approximated by

$$\ell n|S_\tau(f)| \cong \ell n|H_\tau(f)| + \ell n|E_\tau(f) * W(f,\tau)| \qquad (3\text{-}11)$$

In the case of voiced excitation, both $E_\tau(f)$ and $\ell n|E_\tau(f) * W(f,\tau)|$ are periodic in frequency with period $1/T_o$ where T_o is the period of the voicing. Thus the cepstrum

$$c_\tau(t) = F^{-1}\{\ell n|S_\tau(f)|\} \cong F^{-1}\{\ell n|H_\tau(f)|\} + F^{-1}\{\ell n|E_\tau(f) * W(f,\tau)|\} \qquad (3\text{-}12)$$

(the operator $F^{-1}\{\cdot\}$ denotes inverse Fourier transform) has a component made up of impulses at $t = \pm nT_o$, $n = 0, 1, 2, 3, \ldots$, due to $\ell n|E_\tau(f) * W(f,\tau)|$ and another component due to $\ell n|H_\tau(f)|$. If $\ell n|H_\tau(f)|$ is smooth, i.e., a slowly varying function of frequency relative to $\ell n|E_\tau(f) * W(f,\tau)|$, then deconvolution can be achieved by retaining only that portion of $c_\tau(t)$ for which $|t| < T_o$. Note that gating $c_\tau(t)$ in this manner removes the influence of the excitation and the window function except for an impulse at $t = 0$. The effect of this impulse appears as a gain factor on $\hat{h}_\tau(t)$.

For unvoiced excitation, if $e_\tau(t)$ is assumed to have a flat spectrum then $\ell n|E_\tau(f) * W(f,\tau)|$ is essentially constant and its influence on $c_\tau(t)$ is an impulse at $t = 0$. Thus low-time gating of $c_\tau(t)$ preserves the contribution due to $\ell n|H_\tau(f)|$ if the gate is not too short.

Loss of the frequency resolution in this type of deconvolution results from the fact that the influence of $\ell n|H_\tau(f)|$ is not strictly limited to some low-time region of $c_\tau(t)$. Hence truncation of $c_\tau(t)$ causes some loss of detail in frequency for $\ell n|H_\tau(f)|$. This loss of detail

increases as the truncated cepstrum is made shorter.

To the extent that the approximation of Equation (3-10) is valid and the influence of $h_\tau(t)$ on $c_\tau(t)$ is essentially low-time in nature, deconvolution can be accomplished by truncation of the cepstrum with frequency resolution that depends only on the length of the cepstrum truncation. Figures 15 through 22 show computed vocal tract spectra for both voiced and unvoiced speech for three window durations and four cepstrum truncations. Note that window duration has little effect on the frequency resolution when compared to the effect of the cepstrum truncation. Note also that the 12.8 ms window (and probably the 25.6 ms window) seems to violate the assumptions of Equation (3-10) since impulse response durations are typically 10 to 15 ms. However, good deconvolution is obtained as can be seen in Figures 15 through 22. Hammett also reported this result [33].

The nonlinear nature of the cepstral deconvolution process makes a precise definition and analysis of frequency resolution unwieldy. Examination of Figures 14 through 18 shows that a rough measure of frequency resolution can be obtained by determining the increase in bandwidth of the peaks in the vocal tract spectrum as a function of cepstrum truncation length. Measurements made on Figure 14 show that the bandwidth of the main peak is about 100 Hz. Table 1 tabulates this approximate measure of frequency resolution based on Figures 15 through 18, for four cepstrum truncation lengths. Cepstrum truncation length has been defined as L where $c_\tau(t)$ is retained for $|t| \leq L$.

The preceding discussion shows that, under the proper circumstances, the time and frequency resolution of the cepstrum vocoder can be adjusted easily and independently by adjusting certain of the vocoder parameters.

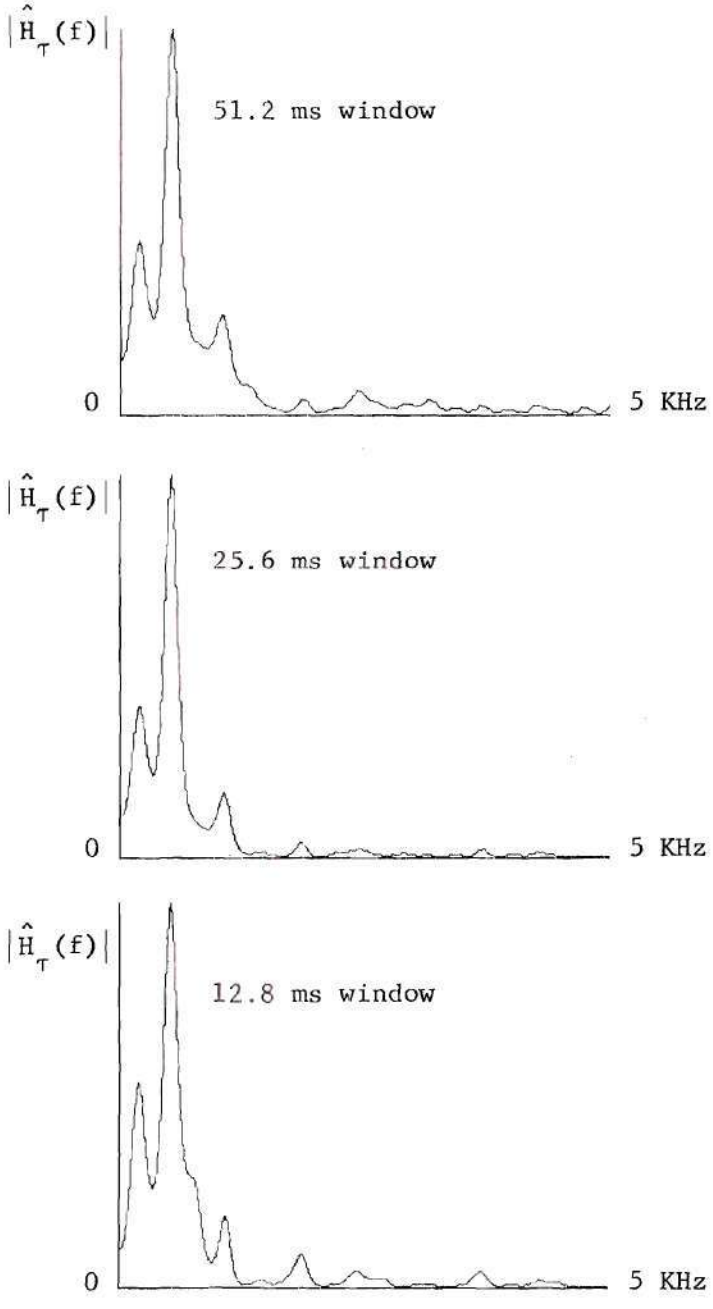


Figure 15. Vocal Tract Spectra for a Sustained /Λ/ Computed with a 4 ms Cepstrum Truncation

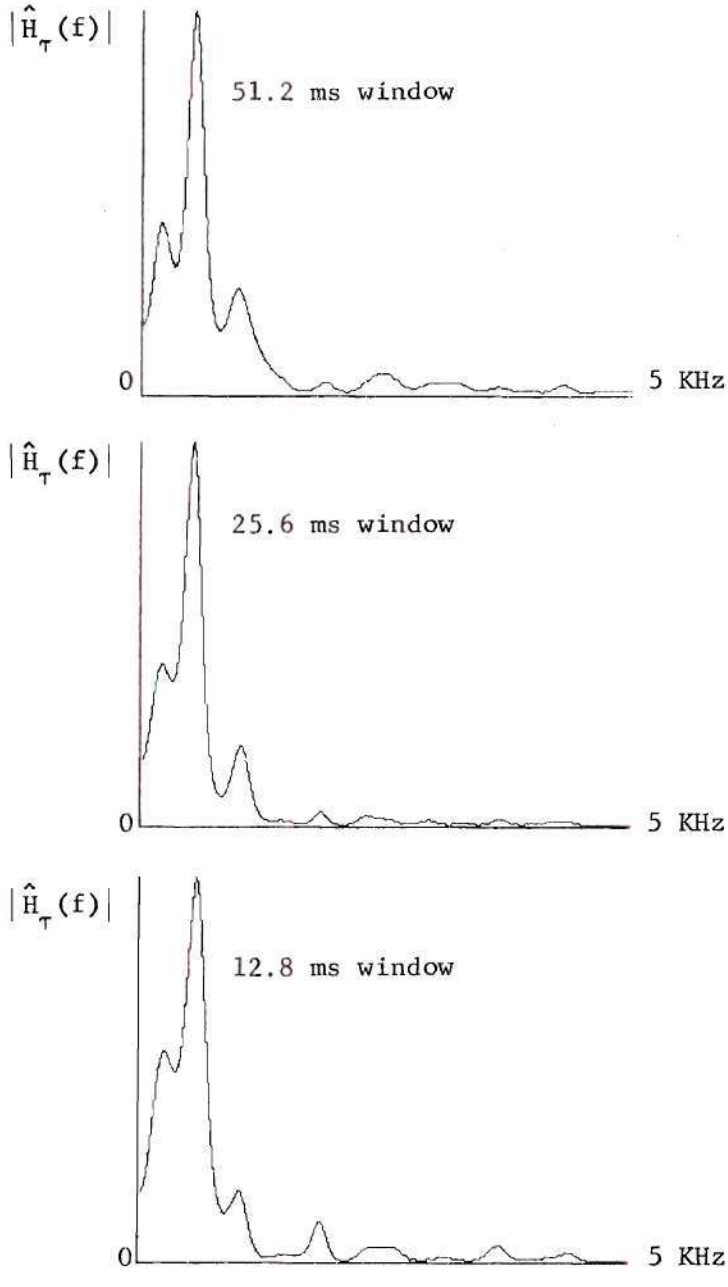


Figure 16. Vocal Tract Spectra for a Sustained /ʌ/ Computed with a 3 ms Cepstrum Truncation

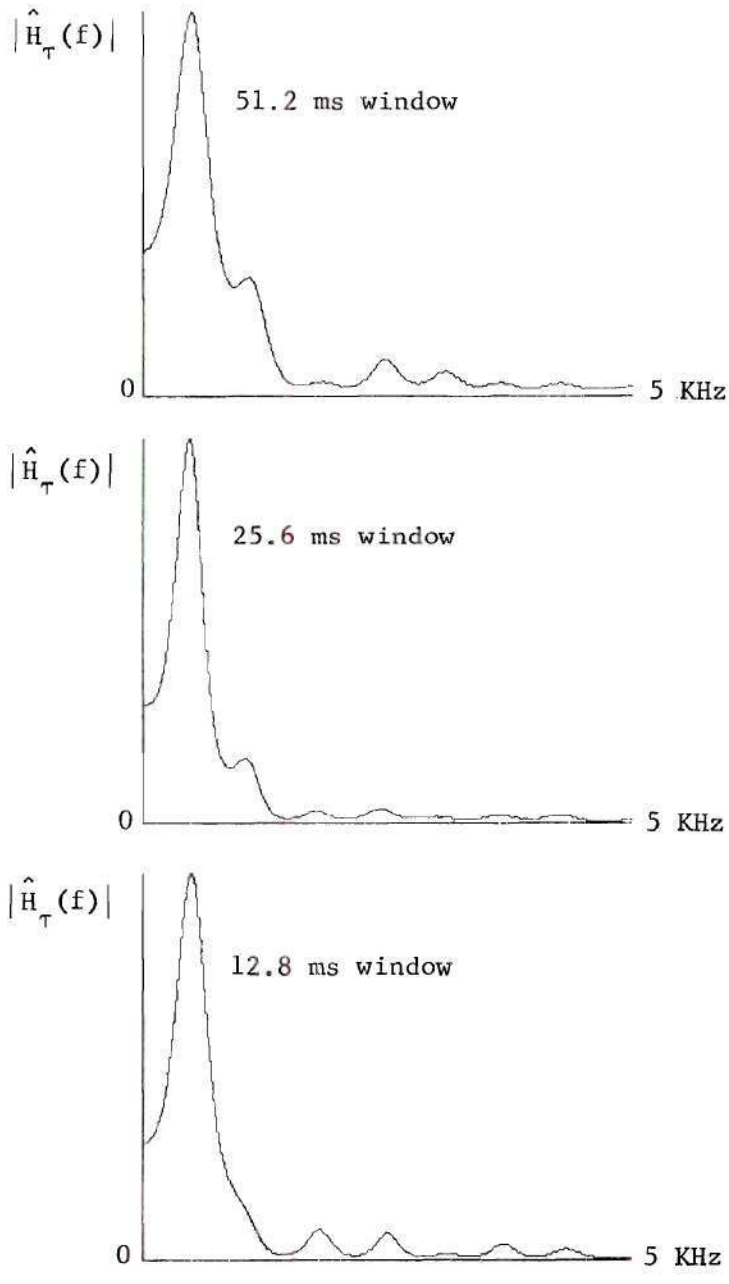


Figure 17. Vocal Tract Spectra for a Sustained /ʌ/ Computed with a 2 ms Cepstrum Truncation

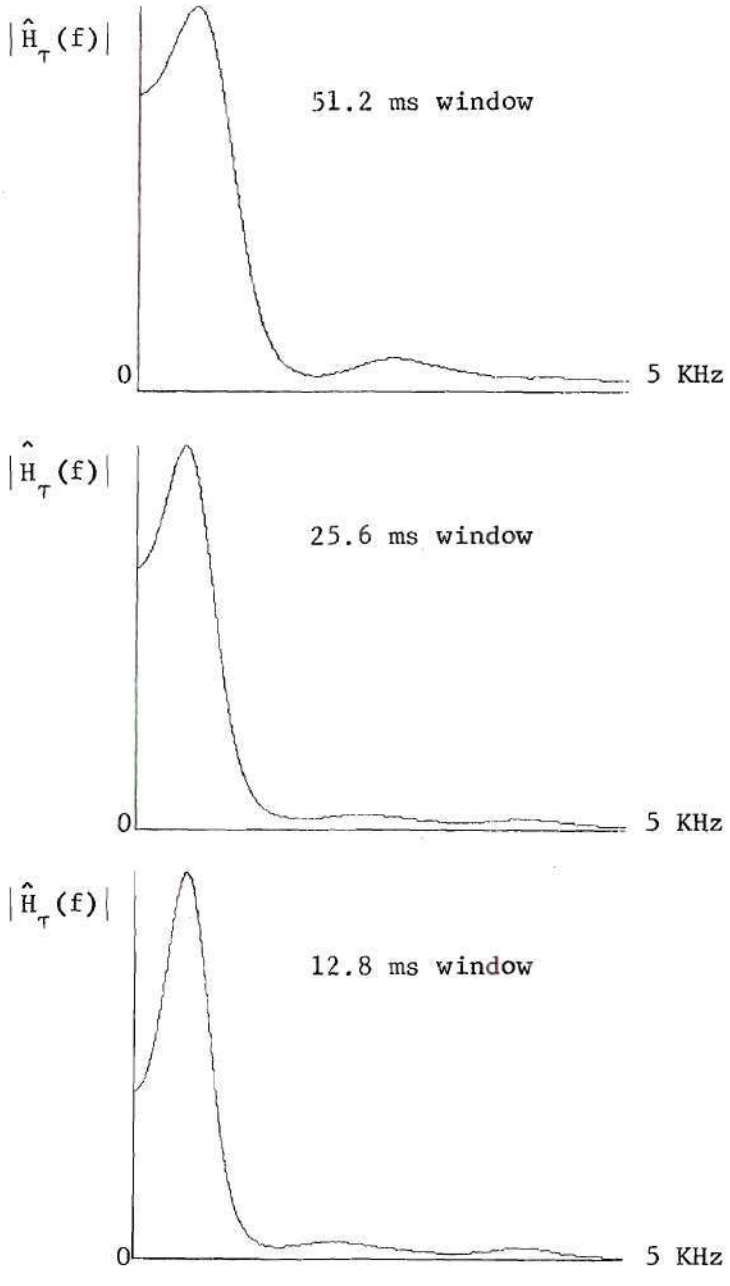


Figure 18. Vocal Tract Spectra for a Sustained /ʌ/ Computed with a 1 ms Cepstrum Truncation

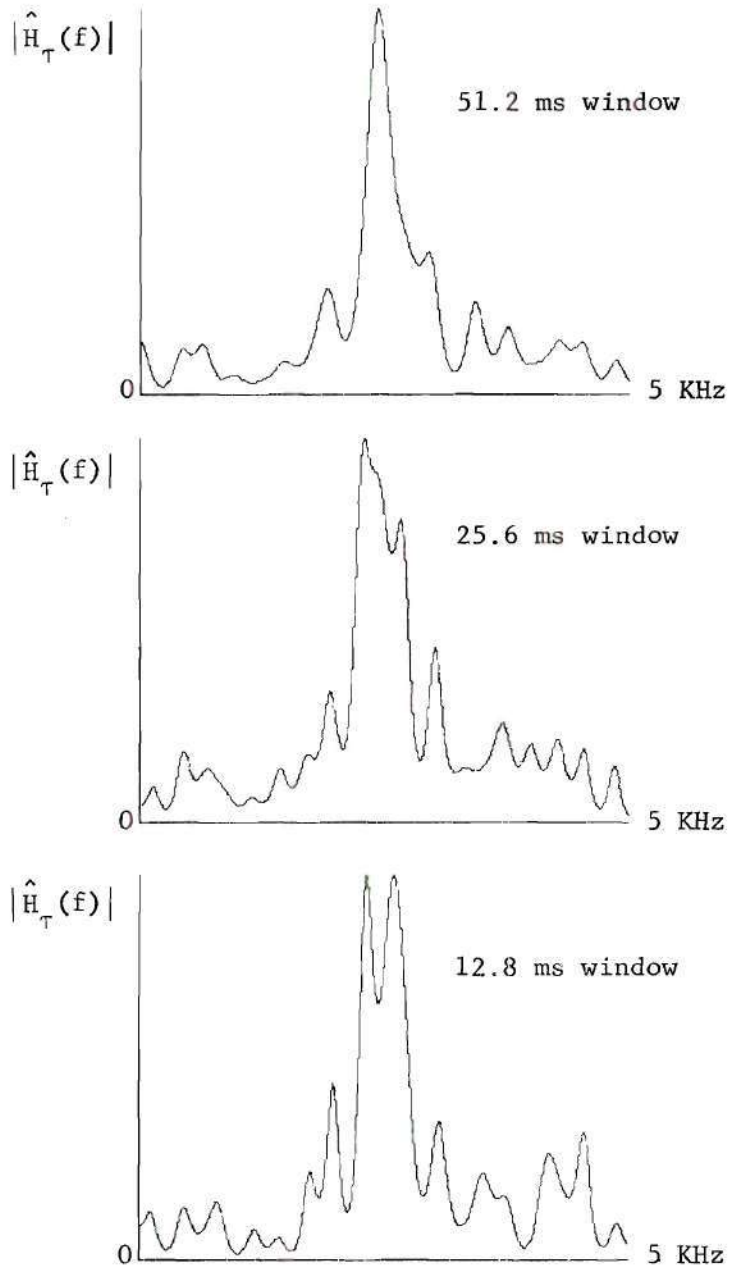


Figure 19. Vocal Tract Spectra for a Sustained /ʃ/ Computed with a 4 ms Cepstrum Truncation

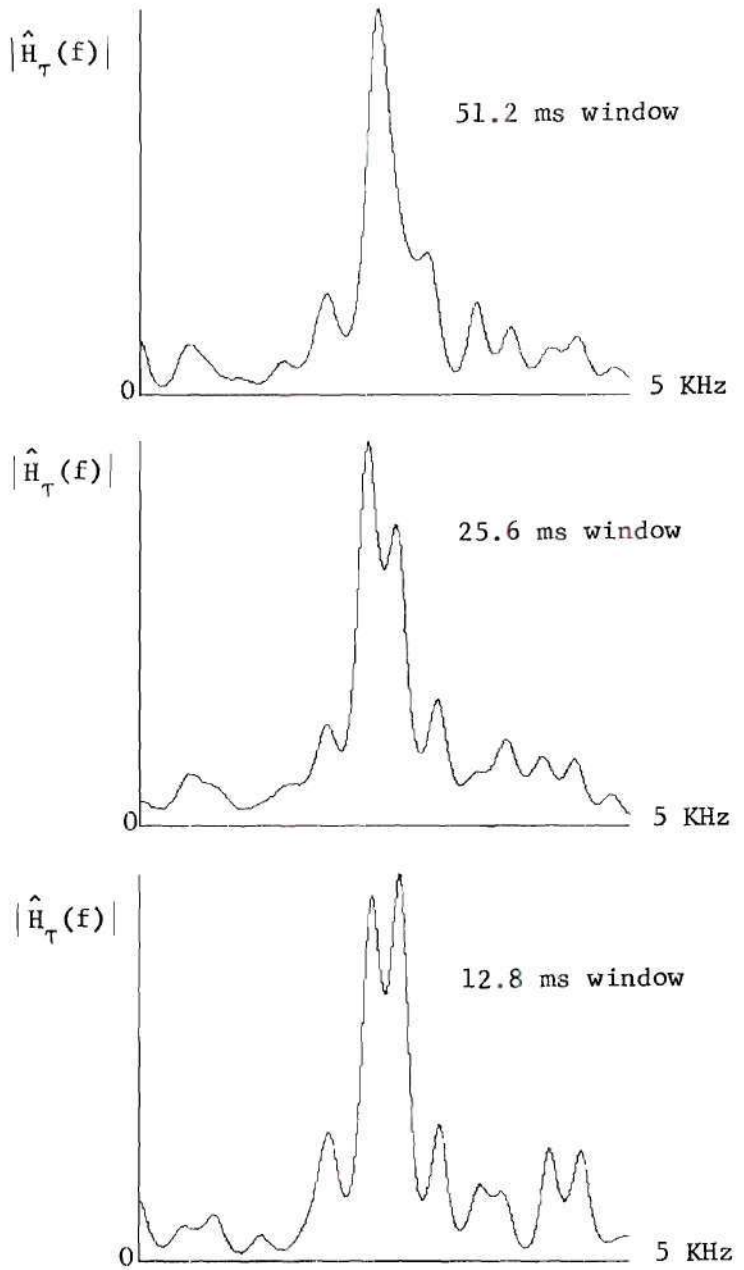


Figure 20. Vocal Tract Spectra of a Sustained /ʃ/ Computed with a 3 ms Cepstrum Truncation

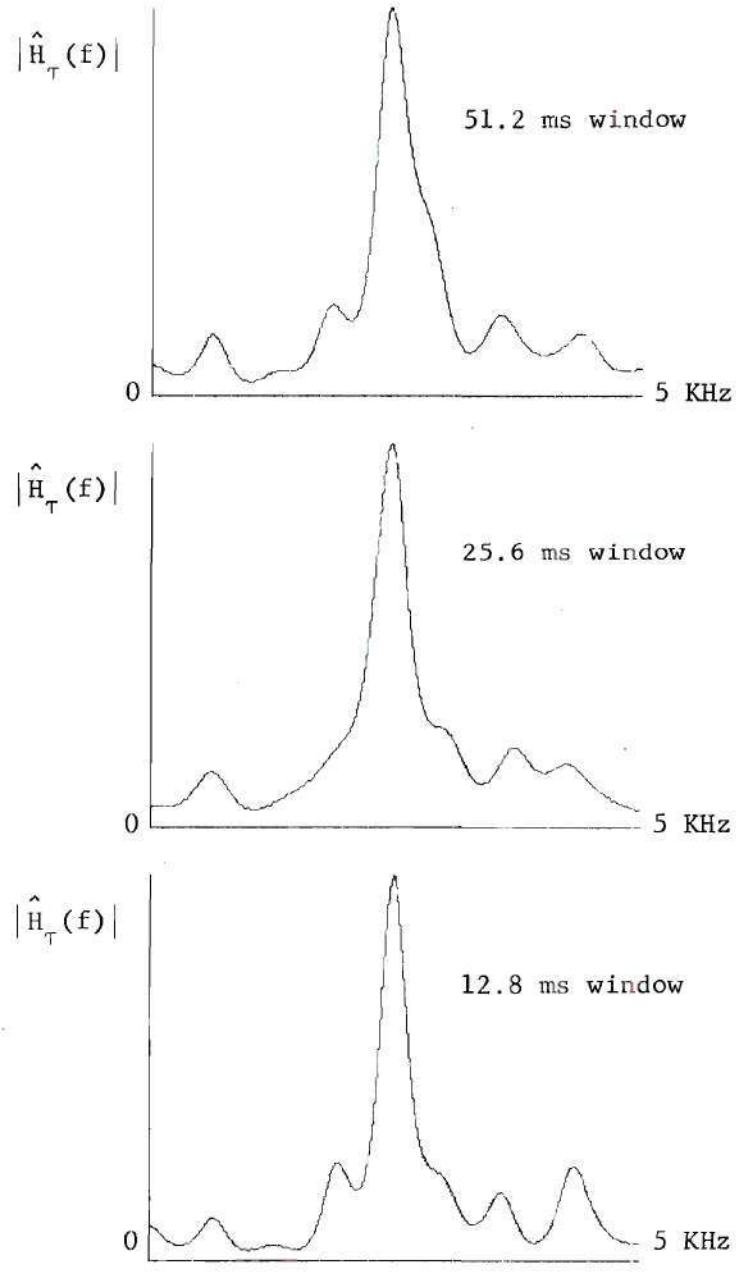


Figure 21. Vocal Tract Spectra for a Sustained /ʃ/ Computed with a 2 ms Cepstrum Truncation

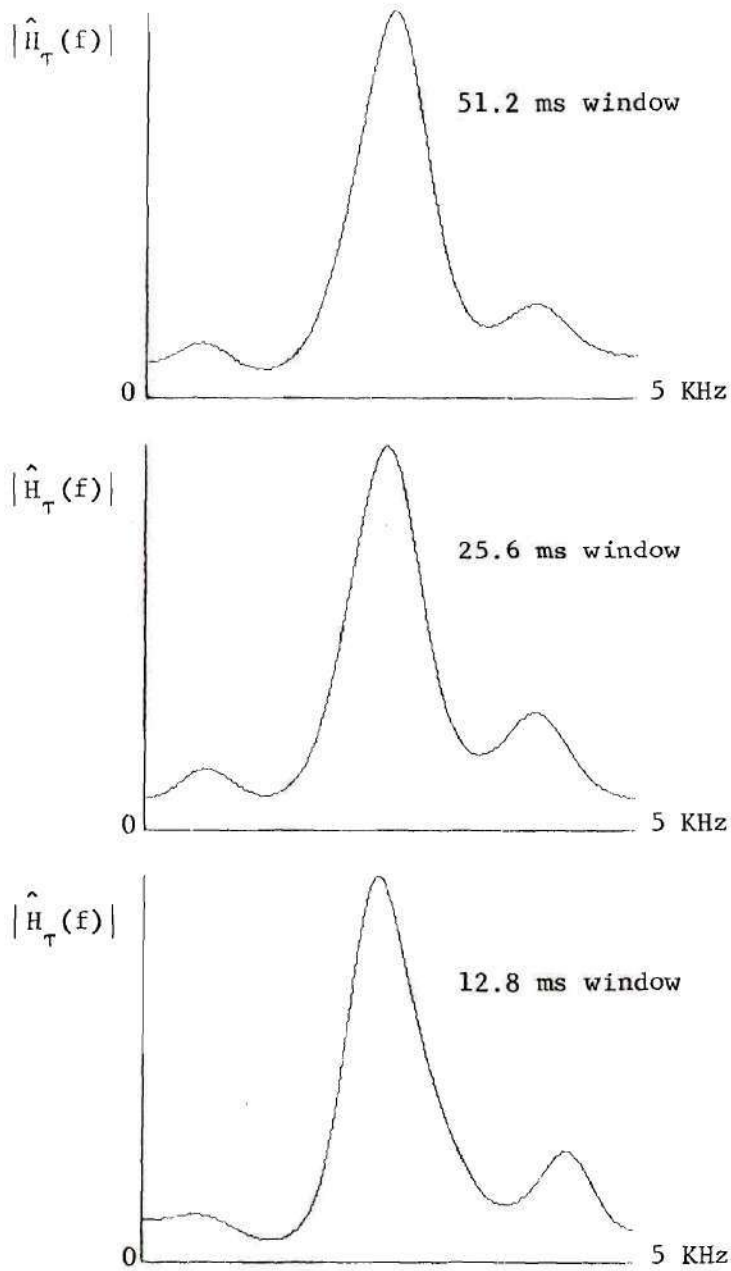


Figure 22. Vocal Tract Spectra for a Sustained /ʃ/ Computed with a 1 ms Cepstrum Truncation

Table 1. Approximate Frequency Resolution of the Cepstrum Vocoder

Cepstrum Length	Frequency Resolution*
4 ms	~ 0 Hz
3	~ 50
2	~ 150
1	~ 400

*Based on the increase in bandwidth of a resonance having a nominal bandwidth of 100 Hz.

This property of the cepstrum vocoder makes it an ideal vehicle for the type of research described in this thesis.

Design Characteristics of the Adaptive Cepstrum Vocoder

The key feature of the vocoder used in this study was its adaptive nature. The adaptive vocoder operated in one of two modes depending on the voiced-unvoiced nature of the input speech. The two modes differed in window duration, frame interval and/or cepstrum truncation length. Mode 1 was used for voiced speech segments and Mode 2 was used during unvoiced and transition segments. The mode was selected in the following way: if the incoming speech was all voiced over the duration of the Mode 1 frame, Mode 1 was chosen. If the incoming speech was not all voiced over the duration of the Mode 1 frame, Mode 2 was chosen. This caused unvoiced segments and segments containing voiced-unvoiced and unvoiced-voiced transitions to be processed in Mode 2.

The vocoder could also be operated in the conventional or non-adaptive mode. Table 2 summarizes the window duration-frame interval combinations and cepstrum truncation lengths employed by the vocoder in both the adaptive and nonadaptive modes. Note that there was very little overlap of frames so that time resolution may be considered equal to the frame interval as stated in the previous section. Approximate frequency resolution may be obtained from Table 1.

Time-frequency resolution as it relates to the specification of the vocal tract impulse response was the object of this study. As a result, it was felt that excitation information should be supplied to the vocoder from an external source. This was done so that excitation

Table 2. Window Durations, Frame Intervals and Cepstrum Lengths Employed by the Simulated Adaptive Cepstrum Vocoder

Mode	Window Duration*	Frame Interval	Cepstrum Length
Adaptive 1	51.2 ms	40 ms	4,3,2 ms
	25.6	20	
Adaptive 2	25.6	20	3,2,1
	12.8	10	
Nonadaptive	51.2	40	4,3,2,1
	25.6	20	
	12.8	10	

*Note that the window duration is uniquely determined by the frame interval.

information of great accuracy could be used and so that the accuracy would not be a function of the vocoder parameters. Excitation extraction was accomplished by manual examination of the speech waveforms [48]. The excitation information was supplied to both the analyzer and synthesizer and was updated every 10.0 ms.

At the synthesizer, 12.8 ms of the vocal tract impulse response was computed and convolved with a train of unit pulses. For voiced excitation the train was periodic with a period equal to the fundamental period. For unvoiced excitation the pulse train had random polarity and a pulse spacing of 1.0 ms. A gain factor of 0.3 was employed during unvoiced synthesis to approximate the level of the input speech. All overlap of impulse responses from previous excitation pulse occurrences was retained in the convolution. Oppenheim [10] suggests that this overlap can be truncated (to make the convolution computationally simpler) with little loss in quality of the synthetic speech. The cepstrum was weighted to achieve minimum-phase synthesis.

Both Oppenheim [10] and Hammett [33] incorporated interpolation of impulse responses into their vocoder designs. The impulse response used at a particular point in the frame was a weighted average of the response computed for that frame and the responses computed for adjacent frames. This was done in order to reduce the characteristic roughness of the synthetic speech which results from abrupt changes in the impulse responses at frame boundaries. Since this roughness of the processed speech is a direct manifestation of imperfect time resolution, interpolation of impulse responses was not used in this vocoder.

The speech input to the vocoder was pre-emphasized using a single-

zero filter providing 6 db/octave boost above 80 Hz. Pre-emphasis was employed to enhance the high-frequency components of the speech signal. These components are generally small compared to the low-frequency components. The output speech was de-emphasized using the corresponding single-pole filter. Informal listening indicated that little improvement (or degradation) in quality resulted from the pre-emphasis and de-emphasis procedure. The procedure was retained to maintain consistency with other on-going speech research.

Computer Simulation of the Adaptive Cepstrum Vocoder

The adaptive cepstrum vocoder was simulated on the Data General Nova 820 computer facility in the School of Electrical Engineering at the Georgia Institute of Technology. This facility is a highly interactive mini-computer system assembled specifically for speech research. The system is built around a Data General Nova 820 16-bit mini-computer having an 800 ns cycle time and 24,576 words of 16-bit core memory. High speed bulk data storage is provided by two Data General moving head disk units with a combined capacity of 2.5 million 16-bit words. A Data General digital cassette recorder provides additional data storage.

Numerous peripheral devices round out the computer system, including a Tektronix 4010 graphics terminal/console, a refresh graphics oscilloscope, a Calcomp Model 565 incremental plotter, a Printec Model 100 line printer, a programmable clock and several analog-to-digital (A/D) and digital-to-analog (D/A) converters.

The vocoder operations were simulated by a collection of Fortran IV programs and subroutines. Appendix A gives a list of these programs and

subroutines and a brief description of each. A variety of system programs were used for A/D and D/A control, plotting and excitation extraction.

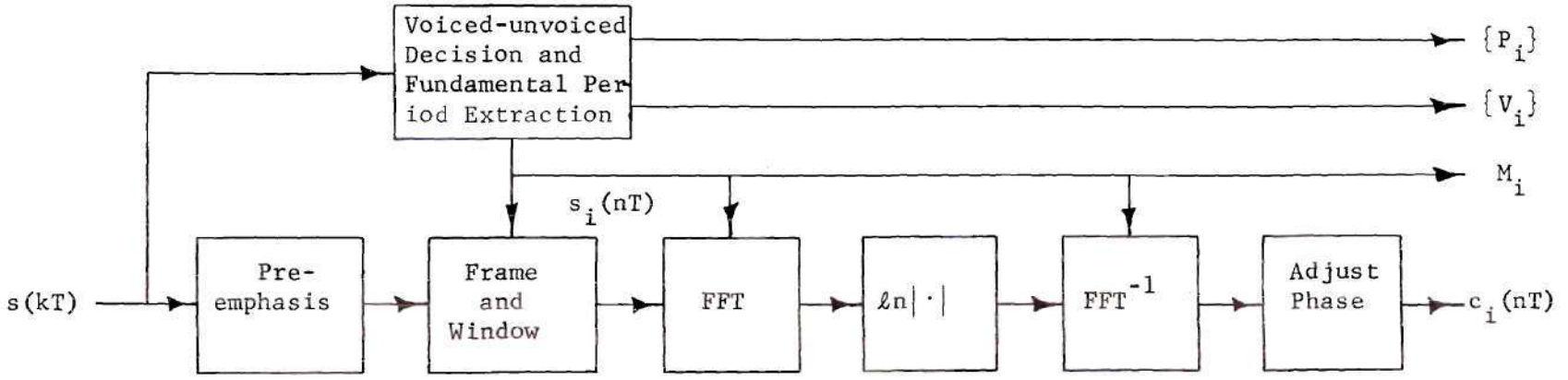
Source speech was played from an analog tape on an Ampex Model 351 tape deck. The tape output was lowpass filtered to 4.0 KHz (24 db/octave attenuation above 4.0 KHz) with a Krohn-Hite Model 3202 filter. The filtered speech was then sampled by a 10-bit A/D converter at a sampling rate of 10.0 KHz. The 10.0 KHz rate was used to compensate for nonideal lowpass filtering. The sampled speech was pre-emphasized using a simulated digital filter and stored on disk in integer form along with manually obtained excitation information.

Block diagrams of the simulated vocoder are shown in Figures 23 and 24 with sampled data notation to emphasize the digital nature of the simulation. A block diagram of the simulation system is given in Figure 25.

During analyzer operation, the sampled speech and excitation information were read from disk. The speech samples were converted to 32-bit floating point representation and the analyzer operations were carried out in floating point arithmetic. The output of the analyzer consisted of floating point cepstrum samples and coded mode information stored on disk.

The synthesizer read the cepstrum, mode, and excitation information from disk and produced floating point synthetic speech which was temporarily stored on disk until it could be quantized to 16-bit integer form and returned to disk.

The final step in the simulation was de-emphasis by a digital filter and digital-to-analog conversion by a 16-bit D/A. The D/A output was lowpass filtered (24 db/octave attenuation above 4.0 KHz) by the Krohn-Hite Model 3202 and recorded on Scotch 206 or 207 tape with the Ampex



s(kT): Input speech samples

$s_i(nT)$: Windowed speech samples for i^{th} frame

$c_i(nT)$: Cepstrum samples for i^{th} frame

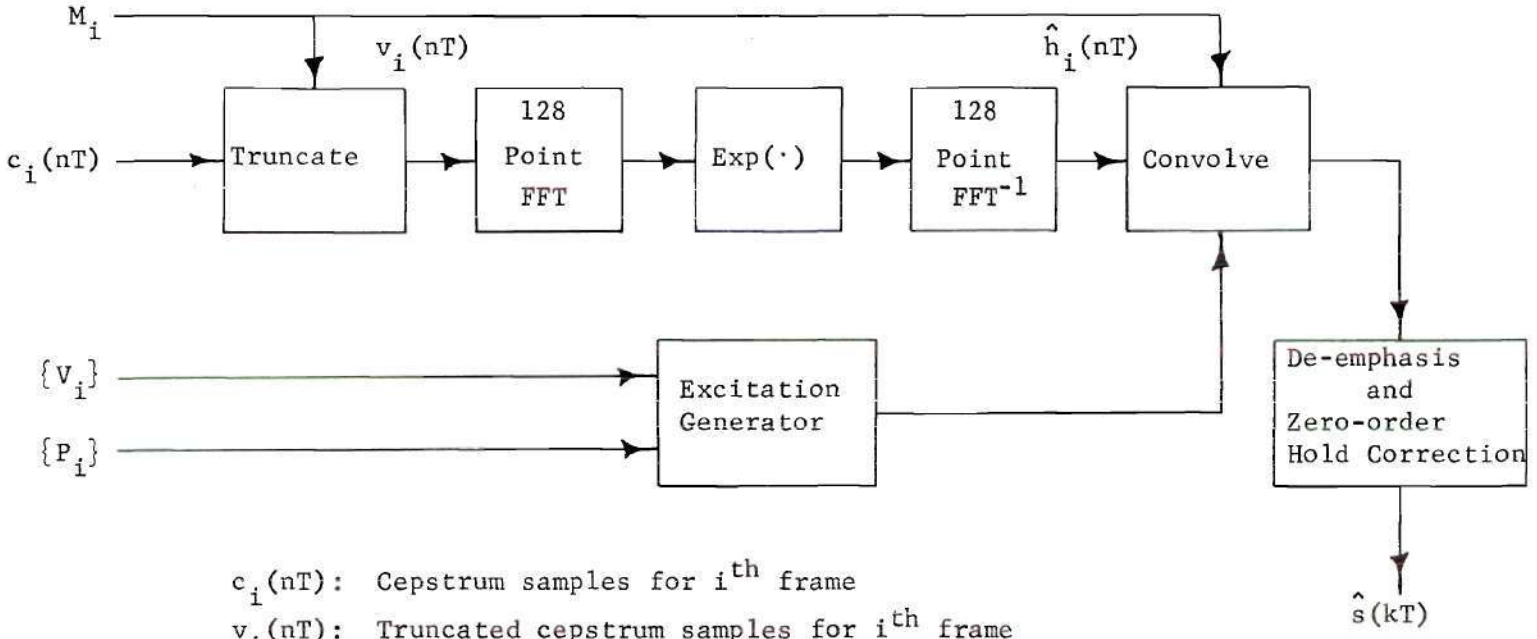
M_i: Mode for i^{th} frame

$\{P_i\}$: Fundamental periods for i^{th} frame

$\{V_i\}$: Voicing decisions for i^{th} frame

T: Sample Spacing

Figure 23. The Simulated Adaptive Cepstrum Analyzer



$c_i(nT)$: Cepstrum samples for i^{th} frame
$v_i(nT)$: Truncated cepstrum samples for i^{th} frame
$\hat{h}_i(nT)$: Synthetic impulse response samples for i^{th} frame
$\hat{s}(kT)$: Synthetic speech samples
M_i: Mode for i^{th} frame
$\{P_i\}$: Fundamental periods for i^{th} frame
$\{V_i\}$: Voicing decisions for i^{th} frame
T: Sample spacing

Figure 24. The Simulated Adaptive Cepstrum Synthesizer

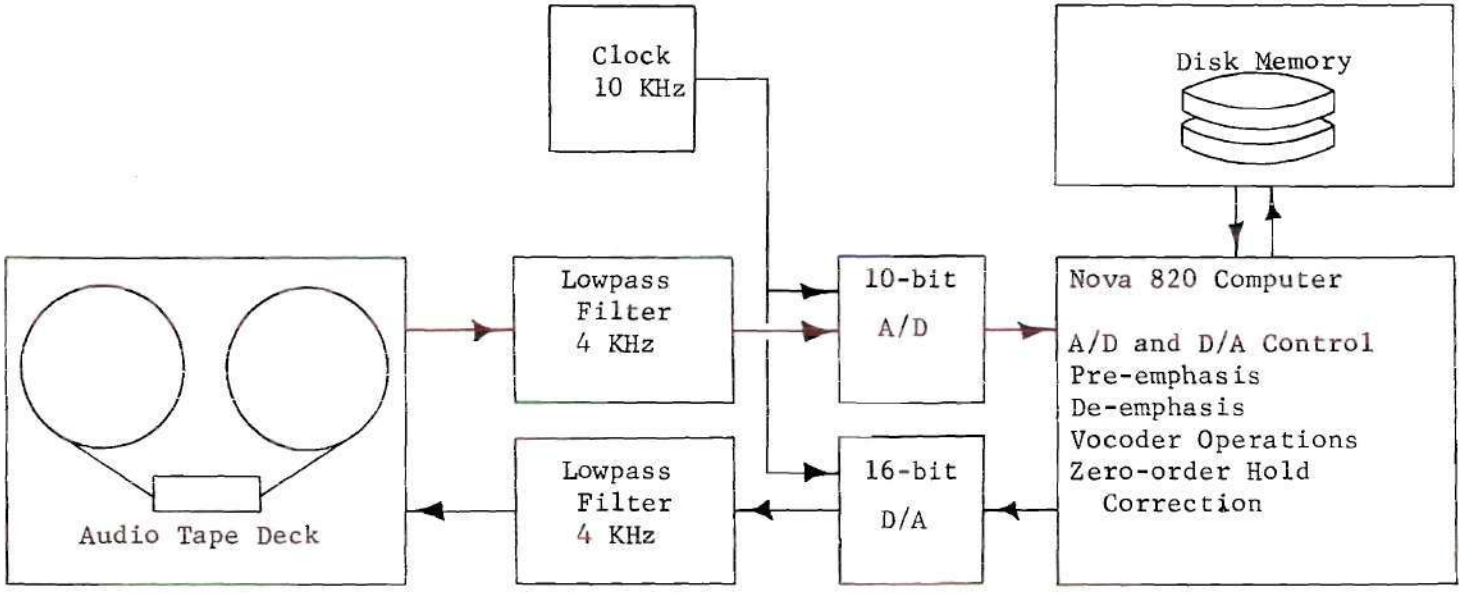


Figure 25. The Vocoder Simulation System

Model 351 tape deck. Compensation for zero-order hold distortion in the D/A converter was provided by a digital filter prior to conversion [35].

Note in Figure 24 that cepstrum truncation was done at the synthesizer rather than at the analyzer. This was done in the interest of simulation speed since the analyzer need be simulated only once for a particular window and frame configuration if truncation is performed at the synthesizer. FFT's of 128 points were used in the synthesizer for the sake of speed. This was possible because only 12.8 ms of the impulse response was used in the convolution.

The simulation of the vocoder was far from real-time since complete processing (analysis and synthesis) of three seconds of speech required approximately one hour of computation time. Thus the advantage of one analyzer run for several synthesizer runs is obvious. Initial development of the vocoder software was carried out with full advantage taken of the computer's interactive capabilities. Once development was completed, the data runs were made in a batch fashion with the synthetic speech being stored on disk for later retrieval and recording. This operating procedure minimized the inconvenience of the long computation time required for the simulation.

The source speech material for this study came from an analog tape supplied by the Department of Electrical Engineering of North Carolina State University. The tape was a duplicate of an original recording of lists 31 through 34 of the 1965 Revised Phonetically Balanced Harvard Sentences [36]. Each list was recorded by two speakers, one male and the other female. The recording was made in a quiet chamber using a Crown Model 822 tape recorder, Scotch 202 tape and an Altec microphone having a

flat frequency response over the speech band [37].

The source sentences used in the simulation were as follows:

1. Men think and plan and sometimes act. (List 32, No. 10)
2. Slide the box into that empty space. (List 31, No. 1)
3. It is late morning on the old wall clock. (List 31, No. 7)
4. Screw the round cap on as tight as needed. (List 32, No. 7)

Sentences 1 and 2 were spoken by the male while sentences 3 and 4 were spoken by the female.

Each of the four source sentences was processed through the simulated vocoder using all possible combinations of the parameters shown in Table 2 for both adaptive and nonadaptive operation. There were 46 different adaptive and nonadaptive vocoder configurations corresponding to the 46 distinct parameter combinations. The 46 configurations are specified in Appendix B. Processing the four sentences through the 46 vocoder configurations resulted in 184 synthetic or processed sentences to be evaluated with respect to quality. For the sake of comparison, the four sampled versions of the source sentences were D/A converted, lowpass filtered at 4.0 KHz and recorded with the synthetic sentences. This corresponded to a 100,000 bps pulse code modulation (PCM) transmission of the signal and was regarded as perfect quality. The addition of these four sentences raised the total number to be evaluated to 188. The method of subjective evaluation used is the subject of the next section.

Subjective Evaluation of the Processed Speech

The subjective evaluation of speech quality is an area of research in its own right. Many techniques for quality measurement have been advanced but as yet no single technique appears to be dramatically better

than the others over a wide range of applications.

One of the chief difficulties in quality or preference measurement arises from the multidimensional nature of the speech perception process. Two different speech samples that are judged to be of equal quality may be judged along different dimensions [38]. Since the dimensions of quality perception are not well defined, most of the measurement methods currently in use attempt to map quality judgments onto a one-dimensional continuum.

An engineering practice for the measurement of speech quality has been prepared by the IEEE Subcommittee on Subjective Measurements [36]. This practice suggests that no single method should be recommended but concludes that three "utilitarian" methods seem appropriate for most engineering requirements. These methods are the Isopreference Method, the Relative Preference Method and the Category Judgment Method. Pachl, Urbanek, and Rothauser [39,40] have made comparative studies of these three methods.

The Isopreference Method makes use of paired comparisons of the speech test signal with high quality speech that is corrupted by noise of some type. The listener is asked to select the noise setting that makes the two signals equal in quality or isopreferent. The measure of quality is expressed in terms of the signal-to-noise ratio of the isopreferent reference signal. The Isopreference Method provides a rank ordering of test signals according to signal-to-noise ratios. However, it is not clear that signal-to-noise ratio is a relevant measure in a vocoder environment.

The Relative Preference Method is similar to the Isopreference Method in that paired comparisons of the test signal to a reference are used. In this case the reference signals represent certain common types

of distortion. The reference signals are first rank ordered and then the order of the test signals is established by comparison to the references.

The Category Judgment Method requires that the listener assign test samples to one of several categories according to their assessment of the signal quality. Typically the categories are assigned numerical values such as 1, 2, 3, 4, 5, and subjective adjectives to describe the quality such as "unsatisfactory," "poor," "fair," "good," and "excellent." Test signals can be rank ordered according to the average score or Mean Category Judgment (MCJ) across listeners. The categories may be established by a process called anchoring in which samples of speech are presented to the listeners and the proper response indicated. Any or all of the categories may be anchored but typically the two extremes are specified. This is done to prevent the error of central tendency or the tendency of listeners to rate all samples near the center of the scale.

Stroh and O'Neal [37] suggest that the selection of a speech quality measurement technique for a particular application should be based on considerations of simplicity, relevance, reliability, and validity. Simplicity refers to simplicity in the construction, administration, and evaluation of the listening test. Relevance deals with the nature of the listener's task in relation to speech communication. Reliability is a measure of the repeatability of the test results and the degree to which different listeners rate a given signal in the same way. Validity refers to the utility of the test in terms of "how well does it measure what I want to measure?"

Based on these considerations, Stroh and O'Neal [37] developed a modified Category Judgment Method which they called the Speech 'Goodness'

Rating Scale. They report good results using this technique in the evaluation of a large group of processed speech signals. Their success with this technique in an application similar to the one at hand motivated the adoption of the Category Judgment Method for this study.

The Category Judgment Method should certainly offer greater simplicity than the other two recommended methods since it requires no production or adjustment of reference signals and minimizes the total number of speech samples that must be presented to the listeners. Simplicity is an especially important consideration when a large collection of test signals is to be evaluated as was the case in this research.

In everyday communication, close comparison between two systems is seldom if ever required. It would seem that the Category Judgment Method more nearly reflects "real world" conditions than either the Isopreference Method or the Relative Preference Method.

Nunnally [41] describes an empirical measure of test reliability based on correlation of standard or z-scores. When applied to quality tests interlistener reliability is defined as the average correlation coefficient between the standardized ratings of pairs of listeners. Overall test reliability can be computed from the average interlistener reliability inserted in the Spearman-Brown formula [37,41]

$$r = \frac{nr_i}{1 + (n - 1)r_i} \tag{3-13}$$

where r is the overall reliability, r_i is the interlistener reliability, and n is the total number of listeners. Accordingly, the Nunnally [41] overall test reliability should be greater than 0.8 for basic research

and greater than 0.9 for applied research. Stroh and O'Neal [37] achieved high reliability ($r = 0.992$ with 25 listeners; $r = 0.996$ with 31 listeners) using their Speech 'Goodness' rating Scale.

The question of validity of any quality measurement method is a difficult one. However, since the Category Judgment Method in essence asks the listener "how well do you like this speech as a means of communication?", it will be assumed that listener responses form a valid measure of quality. The high reliability of the method bolsters this assumption.

The quality test employed in this research was the Category Judgment Method using the nine-level scale illustrated in Figure 26. The use of nine levels rather than the five suggested by the IEEE Recommended Practice on Speech Quality Measurements was based on the results of Urbanek, Pachl, and Rothauser [39] who found that subjects can generally make finer distinction of quality than the five levels allow.

A listening test tape was made by dubbing the 188 speech samples from the master data tapes onto Scotch 206 tape in random order. Two Ampex Model 351 tape decks were used for this dubbing. Anchor samples for the extreme ends of the scale were recorded after every five test samples. The "excellent" anchors were the 100,000 bps PCM signals and the "unsatisfactory" anchors were the outputs from the nonadaptive vocoder using the 51.2 ms window, 40.0 ms frame interval, and 1.0 ms cepstrum truncation. This processor operated at the lowest data rate of any of the processors and its speech was informally judged to be the poorest in quality. After every five test samples, one "excellent" anchor was presented, followed by one "unsatisfactory" anchor. The anchor samples were selected at random. Approximately six seconds were allowed between test

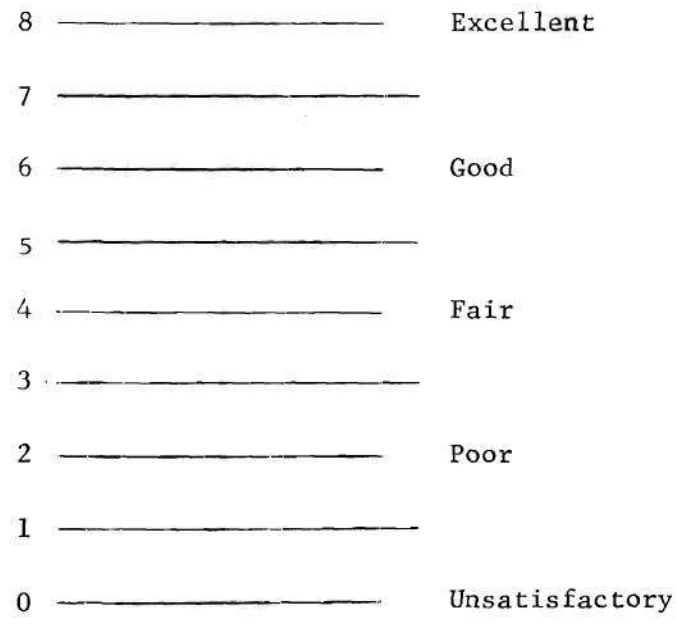


Figure 26. The Category Judgment Scale

signals for the listeners to mark their response on the answer sheet provided.

The test tapes were played on an Ampex 351 tape deck through a Crown Model IC-150 pre-amplifier and a Crown Model DC-300 power amplifier to a network of six pairs of Sennheiser Model HD-414 earphones located in specially constructed listening booths.

A training and familiarization session was held for the listeners a few days prior to the test session. This training period served to acquaint the listeners with the test format and the nature of the test signals. A training tape with the same format as the test tape was used for this session. This training tape was played to the listeners immediately before the actual listening test in order to allow for the adjustment of listening levels and to allow a period for stabilization of listener judgments. Each listener was allowed to adjust the signal level at each ear during this warm-up session and was asked not to change the level during the test. The total time required for the administration of the test was approximately 1 1/2 hours including the warm-up period and several rest periods.

The listening test was administered to 22 listeners in a total of four sessions. The listeners were students, faculty and staff members of the Georgia Institute of Technology. Several of the listeners were familiar with processed speech, but the majority had no previous experience in the area. Two of the listeners reported some hearing loss, but the loss was not evident in their ratings. The ratings of two other listeners were discarded because they diverged significantly from the ratings of the other listeners. One of these listeners showed clear signs of fatigue

or boredom. The other listener reported that he was distracted by details of the test recording that were not related to the research.

Each listener rated each vocoder configuration four times, once for each of the four input sentences. Thus 80 ratings were obtained for each configuration resulting in a total of 3760 ratings.

The average interlistener reliability was computed by correlating the standardized ratings of each possible pair of listeners and taking the average of the correlations. The average interlistener reliability was 0.59 with a corresponding overall test reliability of 0.97 which is above the 0.9 desired for applied research.

MCJ's for each configuration were computed by averaging the 80 ratings for each. A one-way analysis of variance was performed on these MCJ's to determine if there were significant differences among the configurations [42]. Significant differences were detected at the 5% level. A Duncan Multiple Range Test [42] was then performed on the MCJ's to determine which pairs of configurations were not significantly different. The results of this test and the rank ordering of the configurations are shown in Table 3. By convention, any two configurations underscored by the same line are not significantly different at the 5% level. Caution should be exercised in the interpretation of this statistical analysis since such an analysis requires that the Category Judgment Scale be an interval scale. It is not clear that this is the case since it is not clear that the "distances" between the categories are the same.

Appendix B presents a condensation of the listening test results. These results are interpreted and several conclusions are advanced in Chapter IV.

Table 3. Rank Ordering of Configurations and Results of the Duncan Multiple Range Test

Rank	Config. No.	Mode 1 Frame/Cepstrum	Mode 2 Frame/Cepstrum	MCJ	*
1	47	100,000 bps PCM Reference		7.29	
2	30	20/4 ms	10/1 ms	6.15	
3	28	20/4	10/3	6.09	
4	19	20/4	20/3	5.82	
4	45	20/4	20/4	5.82	
5	20	20/4	20/2	5.74	
6	29	20/4	10/2	5.68	
7	21	20/4	20/1	5.37	
8	31	20/3	10/3	4.90	
9	33	20/3	10/1	4.87	
10	32	20/3	10/2	4.82	
11	22	20/3	20/3	4.74	
12	23	20/3	20/2	4.60	
13	42	10/3	10/3	4.45	
14	24	20/3	20/1	4.30	
15	41	10/4	10/4	4.24	
16	10	40/4	10/3	3.87	
17	1	40/4	20/3	3.86	
18	12	40/4	10/1	3.80	
19	35	20/2	10/2	3.76	
20	11	40/4	10/2	3.74	
21	27	20/2	20/1	3.61	
21	34	20/2	10/3	3.61	
22	37	40/4	40/4	3.59	
23	5	40/3	20/2	3.54	
24	15	40/3	10/1	3.52	
25	2	40/4	20/2	3.48	
26	25	20/2	20/3	3.46	
27	14	40/3	10/2	3.44	
28	6	40/3	20/1	3.43	
29	26	20/2	20/2	3.41	
30	13	40/3	10/3	3.37	
31	3	40/4	20/1	3.27	
32	4	40/3	20/3	3.24	
33	43	10/2	10/2	3.18	
34	36	20/2	10/1	3.16	
35	38	40/3	40/3	2.85	
36	7	40/2	20/3	2.80	
37	18	40/2	10/1	2.73	
38	17	40/2	10/2	2.66	
39	8	40/2	20/2	2.60	
40	16	40/2	10/3	2.59	
41	9	40/2	20/1	2.46	
42	39	40/2	40/2	1.65	
43	46	20/1	20/1	1.62	
44	44	10/1	10/1	1.48	
45	40	40/1	40/1	1.15	

*Configurations connected by the same line are not significantly different according to the Duncan Multiple Range Test ($p = 0.05$).

Summary

A program of research employing an adaptive cepstrum vocoder was carried out to study the perceptual effects of reduced time-frequency resolution in vocal tract specification in a vocoder environment. The cepstrum vocoder was selected because of its time-frequency resolution properties. The vocoder was digitally simulated.

The performance of the vocoder under different time-frequency resolution conditions was evaluated subjectively using the Category Judgment Method. Conclusions drawn from the results of the subjective evaluation are presented in Chapter IV.

CHAPTER IV

RESULTS AND CONCLUSIONS

The goal of this dissertation research was to study the effects of vocoder time-frequency resolution on the perceived quality of vocoder speech. Chapter III described the design, simulation, and subjective evaluation of an adaptive cepstrum vocoder. The present chapter sets forth several conclusions concerning time-frequency resolution and speech quality. These conclusions were drawn from the results of the Category Judgment evaluation of the vocoded speech signals obtained from the simulation phase of the research.

It is hoped that these conclusions will contribute to the further understanding of the speech perception process and that they will point the way toward improvements in vocoder design. In addition, these conclusions should suggest areas for further research.

Conclusion One: Adequate Vocoder Time Resolution

Figure 27 shows the MCJ's for the nonadaptive vocoder plotted as a function of frame interval and cepstrum length. For each condition of frequency resolution, the two shorter frames showed an advantage in quality over the 40.0 ms frame. The performances of the 10.0 ms frame and the 20.0 ms frame were comparable for all cepstrum lengths except 4.0 ms where the 20.0 ms frame was significantly better. This difference is unexplained but may be the result of some sort of time-frequency trading or

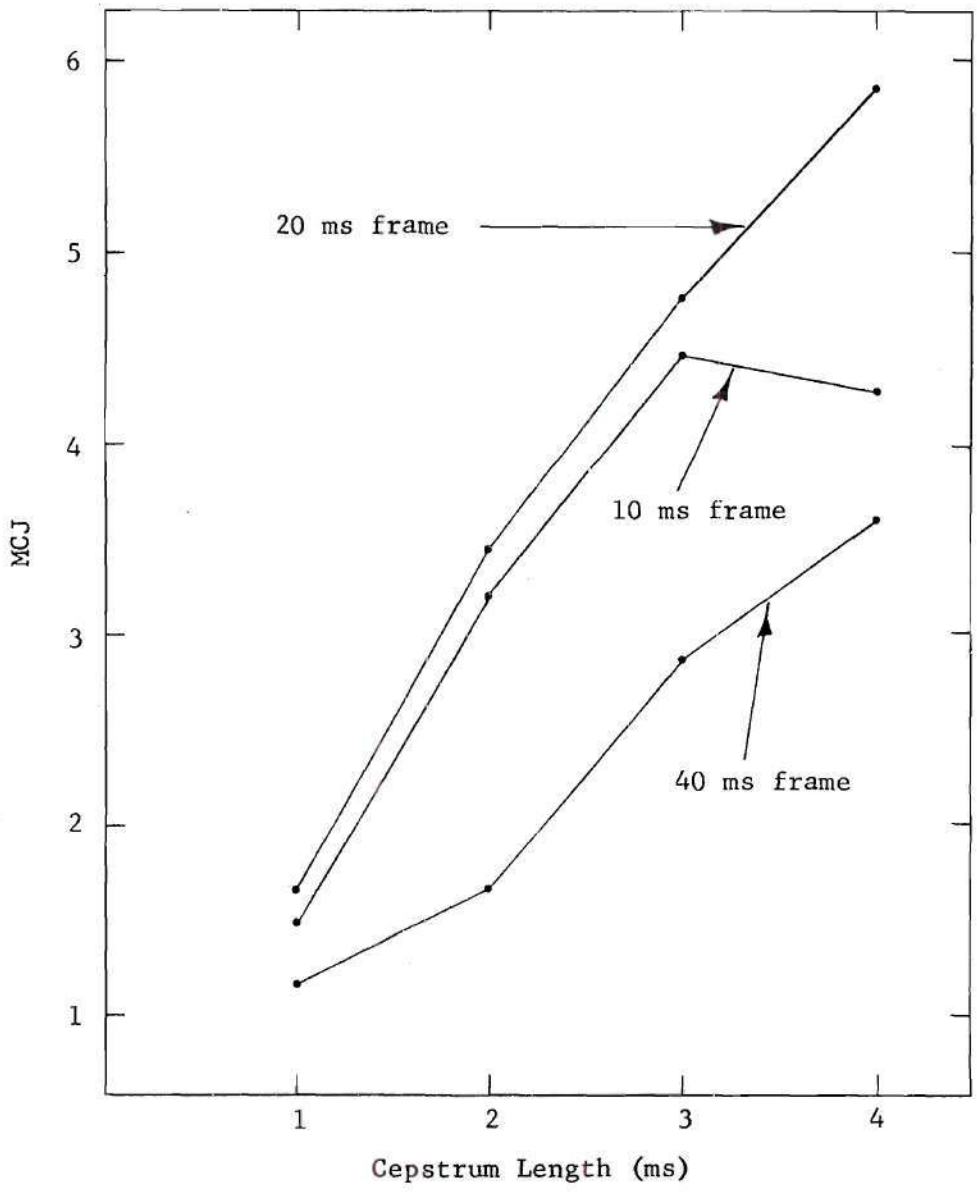


Figure 27. Performance of Nonadaptive Configurations

simply a "quirk" in the data. It should be noted that the quality at a cepstrum length of 1.0 ms was essentially the same for all three frames, suggesting that poor frequency resolution was the predominant factor at that point.

The conclusion to be drawn from this set of results is that maintaining time resolution better than about 20.0 ms seems to provide no improvement in speech quality.

Conclusion Two: Time-Frequency Trading in Quality Perception

An examination of the nonadaptive vocoders with 20.0 ms and 40.0 ms frames in Figure 27 shows that configurations with equivalent data rates (for example, 40.0 ms frame and 4.0 ms cepstrum compared to 20.0 ms frame and 2.0 ms cepstrum) have roughly equivalent quality. This observation may be interpreted as evidence of time-frequency trading in speech perception.

Conclusion Three: The Effect of Reduced Frequency Resolution in Unvoiced and Transition Regions

Figures 28 and 29 show plots of the MCJ's for the various adaptive vocoder configurations. Note that reducing the Mode 2 cepstrum length had no appreciable effect on the speech quality. This result was independent of Mode 1 frame, Mode 1 cepstrum, and Mode 2 frame. The conclusion is that frequency resolution can be reduced considerably in unvoiced regions and regions of voiced-unvoiced or unvoiced-voiced transition with little or no loss in speech quality. This is an important result for the design

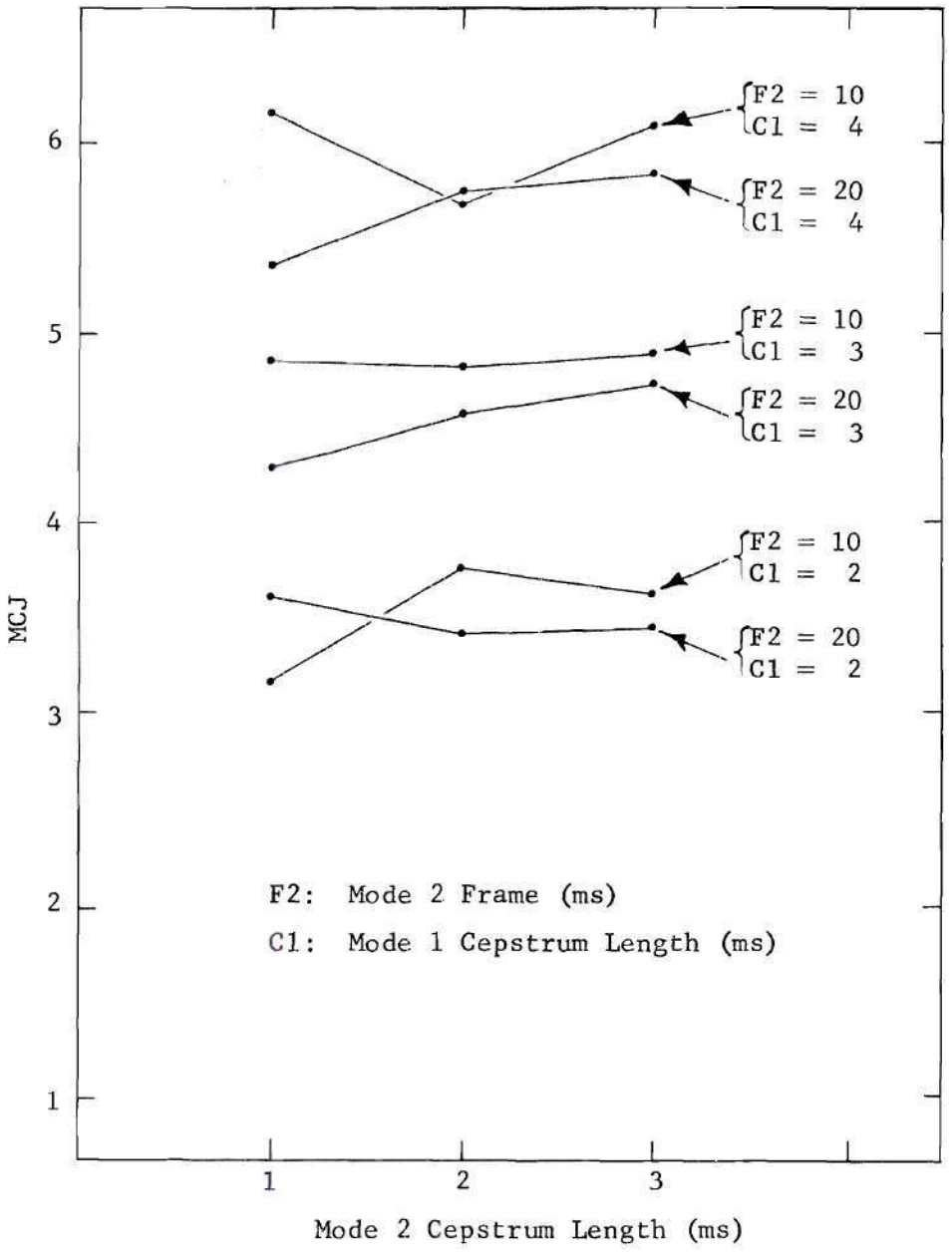


Figure 28. Performance of Adaptive Configurations with 20 ms Mode 1 Frame

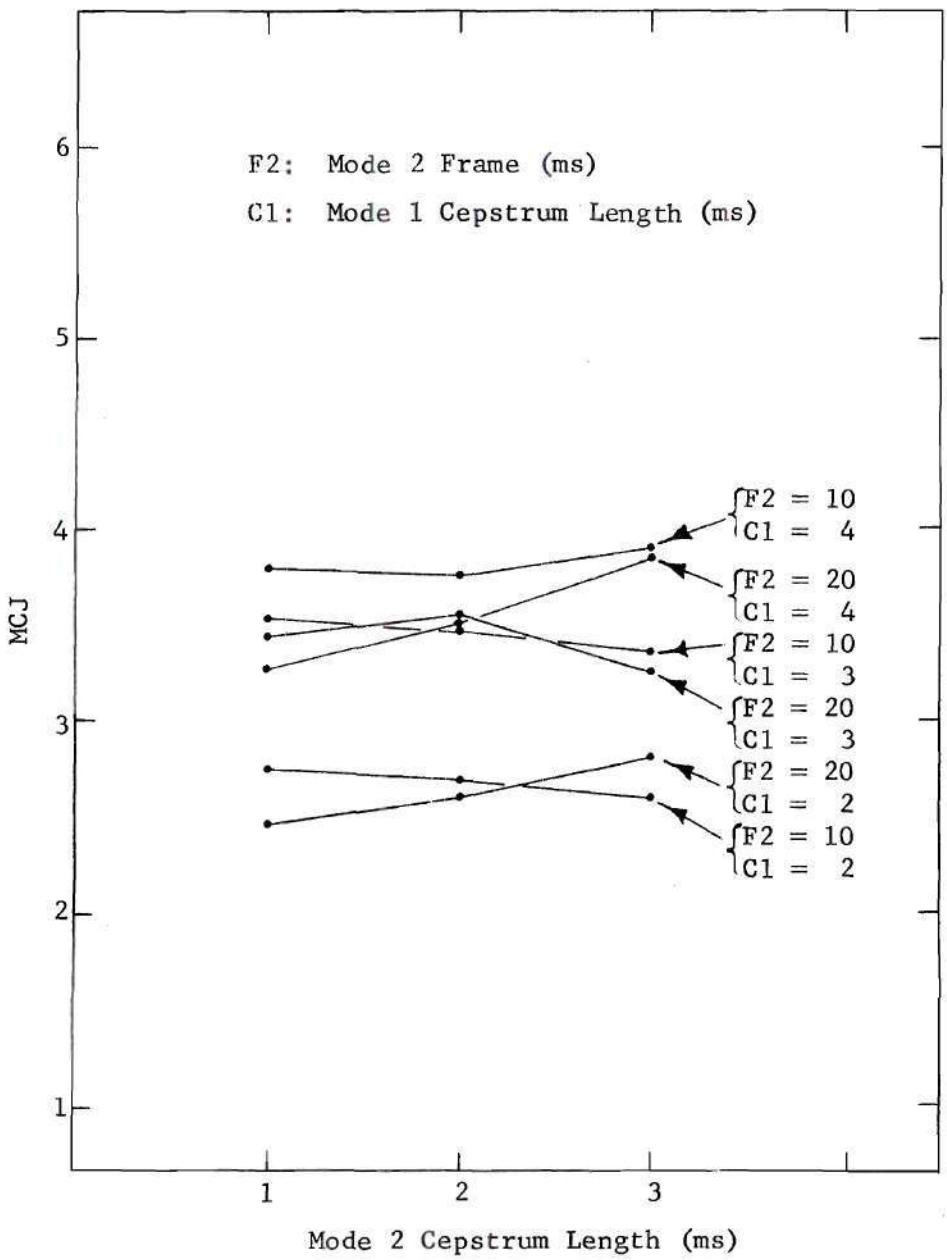


Figure 29. Performance of Adaptive Configurations with 40 ms Mode 1 Frame

of adaptive vocoders which must maintain a constant data rate. This conclusion also lends support to the notion of time-frequency trading in speech perception.

Conclusion Four: The Effect of the Adaptive Strategy

The effect of the adaptive strategy is displayed in Figure 30. This figure gives a comparison of the MCJ's for several versions of the adaptive and nonadaptive vocoders. For the 20.0 ms frame at a fixed cepstrum length, adapting to the 10.0 ms frame for unvoiced and transition regions resulted in no improvement in quality. This observation is in agreement with Conclusion One and suggests that time resolution of about 20.0 ms is sufficient for vocoders.

For the 40.0 ms frame and 4.0 ms cepstrum, adaption had no noticeable effect. However, for the 40.0 ms frame and either the 3.0 ms or 2.0 ms cepstrum, adapting to a shorter frame in unvoiced and transition regions led to improved quality. The improvement obtained was independent of the Mode 2 frame used, giving still more support to Conclusion One. The improvement was roughly equivalent to increasing the cepstrum length of the nonadaptive vocoder by 1.0 ms. It is not clear why adaption produced no improvement in quality for the 4.0 ms cepstrum case.

It appears that 20.0 ms time resolution is adequate for vocoder applications. Thus adaption in a system which normally maintains 20.0 ms resolution or better yields no improvement in performance. For systems that do not normally employ such good time resolution, adaption seems to offer considerable potential.

It should be pointed out that interpolation of impulse responses

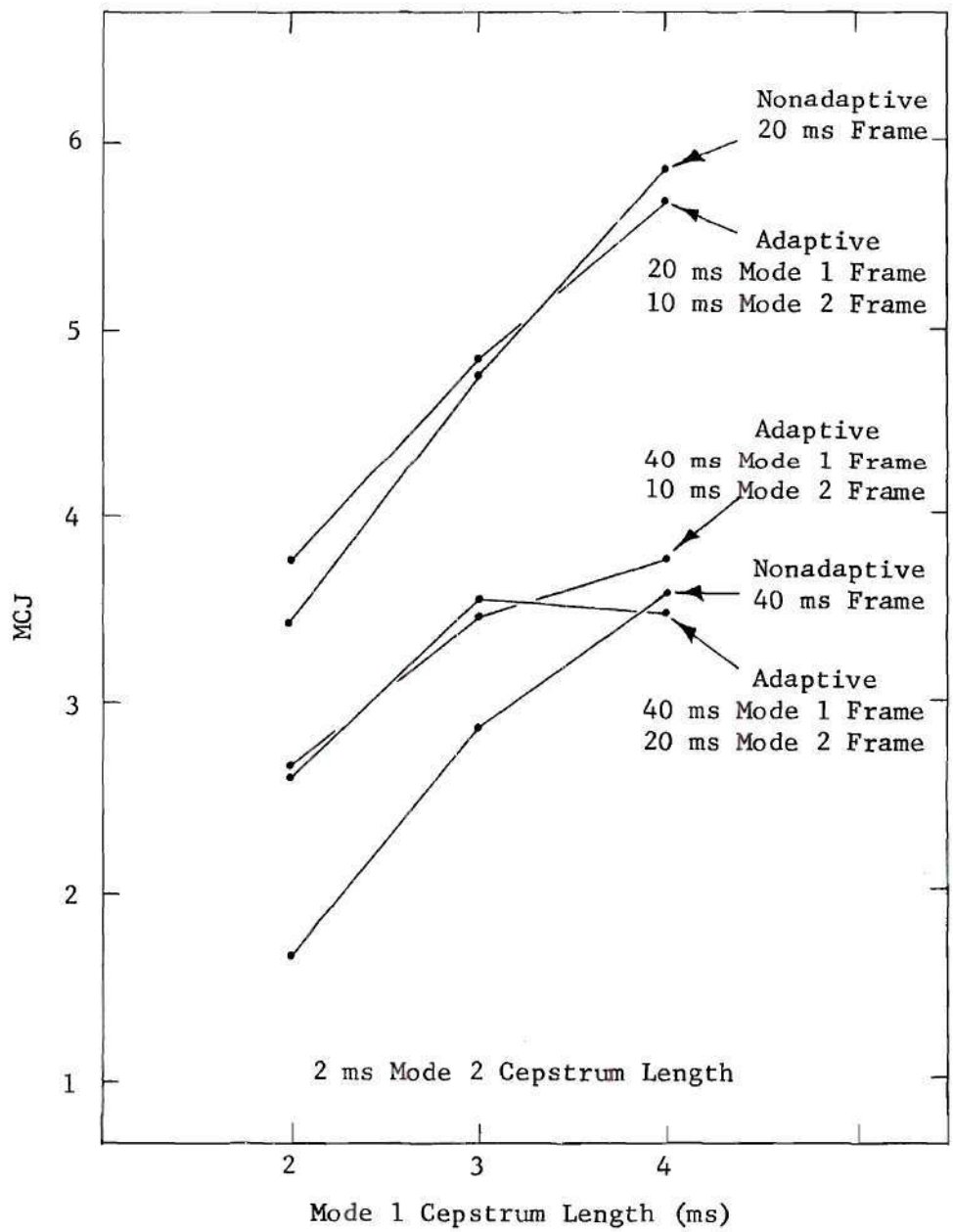


Figure 30. Effect of Adapting to a Shorter Frame in Unvoiced and Transition Regions

would probably improve the performance of the vocoders using 40.0 ms frames in voiced regions. Thus a combination of interpolation and adaption in these systems might well produce good quality speech at quite low data rates.

Discussion

The conclusions presented above should be regarded as tentative for several reasons. Only two speakers and four utterances were used in the research. Only one type of vocoder was employed and speech quality was judged by only one of several methods available. It is certainly conceivable that a similar study with different source material, a different vocoder, or a different quality measurement technique could produce different results.

At an early stage in this research it became evident that the two speakers used produced vocoded speech of widely differing quality. This behavior can be seen in Appendix B. The male speaker rated higher in quality than the female in all instances except three. This dependence of quality on the speakers may have colored the results of this study.

Stroh and O'Neal [37] have reported a similar phenomenon in a study of speech encoding techniques. It must be pointed out, however, that the source material for this work was obtained from Stroh and O'Neal and hence the same two speakers were used in both studies. Melsa, et al. [43] report poorer quality female speech compared to male with a different set of speakers as does Barnwell [44].

If a speaker sex dependent quality difference does exist in vocoder speech, it may be possible to explain it in terms of fundamental

periods. Females typically have shorter fundamental periods than males so that there is more overlap of impulse responses during voiced speech and deconvolution is more difficult. An equivalent explanation can be given in the frequency domain. Since female speakers generally have shorter fundamental periods, the spectral lines in the short-time spectrum are spaced further apart so that the envelope due to the vocal tract is "sampled" less often in frequency making deconvolution by smoothing more difficult. In the particular case of the cepstrum vocoder, short fundamental periods cause the excitation portion of the cepstrum to encroach on the vocal tract portion with a resulting loss of quality in the deconvolution.

Summary

Evaluation of vocoded speech obtained during the simulation phase of this research has led to several tentative conclusions regarding time-frequency resolution and vocoder speech quality.

It appears that time resolution on the order of 20.0 ms is adequate for vocoder applications. Thus systems that maintain this level of time resolution do not benefit from adapting to better time resolution in unvoiced and transition regions. On the other hand, improvements in quality and/or reductions in data rates may be possible if systems that normally employ time resolution worse than about 20.0 ms make use of adaptive resolution.

Evidence that speech perception trades between time and frequency has been obtained by noting that reduction of frequency resolution in unvoiced and transition regions brings about no noticeable degradation

in quality. The time-frequency trading notion is further supported by the observation that systems with the same data rate but different time-frequency resolution conditions may be judged equivalent in quality.

The question of quality dependence on speaker sex has also been raised by this research.

CHAPTER V

RECOMMENDATIONS FOR FURTHER WORK

During the course of this research it has become apparent that additional work is needed in this and related areas. Also, a question not directly related to the time-frequency resolution problem has been raised. As a result, four areas of further research have been formulated. These areas are:

1. A more extensive investigation of the time-frequency resolution question in relation to vocal tract coding.

2. The development of a family of test signals to measure the time-frequency resolution properties of vocoders.

3. A study of the effects of vocoder time-frequency resolution in excitation coding.

4. An investigation of the speaker sex and vocoder quality question.

Extension of the Present Work

The conclusions of this dissertation research are tentative in nature because of the limited scope of the study, i.e., limited with respect to speech source material, vocoder used, and quality measure employed. There is a great need for a more extensive study of time-frequency resolution in vocal tract coding. Such a study should make use of a variety of speakers and source material, several types of vocoders, and

perhaps several types of quality evaluation techniques. The great problem with an investigation of this type is the large amount of processed speech to be subjectively evaluated. The development of an automated testing procedure would facilitate such a study considerably.

Development of Time-Frequency Resolution Test Signals

A precise analysis of the time-frequency resolution properties of many vocoders is a difficult if not impossible task. The evaluation of vocoder performance would be a much easier task if there existed a family of signals to test the time-frequency resolution capabilities of vocoders. If the relation between time-frequency resolution and vocoder quality can be firmly established, the use of time-frequency resolution test signals could provide useful predictions of quality in vocoder design. Of course the design of such a family of test signals would first require the development of acceptable definitions of time and frequency resolution.

Time-Frequency Resolution in Excitation Coding

This dissertation has been concerned only with time-frequency resolution in the specification of the vocal tract impulse response. A similar study of time-frequency resolution in the specification of the vocal tract excitation is needed if vocoder design is to be optimized. A study of this type could be carried out by driving a high quality vocoder with excitation information derived from excitation extractors incorporating a variety of time-frequency resolution conditions. The processed speech output of the vocoder could then be evaluated with respect to quality and the quality related to the time-frequency resolution of the excitation coding.

The Male-Female Vocoder Quality Problem

Perhaps the most intriguing question raised by this research concerns the existence of a difference in vocoded speech quality based on speaker sex. Further research into this problem seems appropriate. The first question to be answered by the research is whether or not such a sex-based difference actually exists. If the difference is found to be real, its source should be determined and consideration given to changes in the speech model and/or speech analysis-synthesis techniques in order to overcome the problem.

APPENDIX A

SOFTWARE FOR SIMULATION OF THE ADAPTIVE CEPSTRUM VOCODER

This appendix gives a list and brief descriptions of the Fortran IV programs and subroutines written for the simulation of the adaptive cepstrum vocoder.

1. XMTR--Main control program for the analyzer operations of the vocoder. The inputs to this program were sampled speech in integer form and excitation information. The outputs were floating point cepstrum samples and mode information.

2. RCVR--Main control program for the synthesizer operations of the vocoder. The inputs to this program were floating point cepstrum samples, mode information, and excitation information. The output was sampled synthetic speech in integer form.

3. PICMOD--Mode decision subroutine. This subroutine examined the voiced-unvoiced decisions for input speech and determined which of the two adaptive modes should be employed in the next frame.

4. GETSEG--Speech input subroutine. This subroutine read integer speech data from disk memory into core memory. The block of data read in was controlled by the window duration and frame interval of the current and previous frames. This routine converted the speech from integer to floating point form.

5. ENCODE--Cepstrum analyzer subroutine. This subroutine performed the operations of the analyzer. Its input was floating point speech and

its output was the cepstrum in floating point form.

6. CEPOUT--Cepstrum output subroutine. This subroutine read floating point cepstrum values from core memory to disk. This subroutine also coded mode information into the cepstrum output.

7. CEPIN--Cepstrum input subroutine. This subroutine read floating point cepstrum samples and mode information from disk to core memory.

8. DECODE--Cepstrum synthesizer subroutine. This subroutine calculated the approximation to the vocal tract impulse response from the input cepstrum values.

9. SYNTH--Convolution subroutine. This subroutine convolved the synthetic vocal tract impulse response with the appropriate train of impulses to produce synthetic speech.

10. RANDOM--Random polarity subroutine. This subroutine generated the random polarities required in the production of unvoiced excitation.

11. SPKOUT--Speech output subroutine. This subroutine read floating point synthetic speech from core to disk for temporary storage. The speech was later read back in, converted to integer form and read out again.

12. LOADHW--Hanning window subroutine. This subroutine loaded a floating point array with samples of the Hanning window function. This was done once per vocoder run so that the window function did not have to be computed for each frame of speech processed.

13. FFT--Fast Fourier Transform subroutine. This subroutine computed discrete Fourier transforms using an FFT algorithm due to Uhrich [45].

14. FFTSIN--FFT sine subroutine. This subroutine loaded an array with samples of the sine function required by the FFT subroutine. This

was done once per vocoder run in the interest of speed.

A variety of other programs and subroutines were used to support the simulation of the adaptive cepstrum vocoder. These programs and routines handled such things as graphics, A/D and D/A conversion, digital filtering, and listening test data analysis.

APPENDIX B

CONDENSED LISTENING TEST RESULTS

Config. No.	Mode 1 Frame/Cepstrum	Mode 2 Frame/Cepstrum	Male MCJ	Female MCJ	Overall MCJ	Std. Dev.*
1	40/4 ms	20/3 ms	4.85	2.87	3.86	1.12
2	40/4	20/2	4.43	2.52	3.48	1.08
3	40/4	20/1	3.85	2.70	3.27	1.15
4	40/3	20/3	4.15	2.32	3.24	0.96
5	40/3	20/2	5.00	2.07	3.54	0.92
6	40/3	20/1	4.25	2.60	3.43	1.09
7	40/2	20/3	3.80	1.80	2.80	1.16
8	40/2	20/2	3.45	1.75	2.60	0.96
9	40/2	20/1	2.93	2.00	2.46	0.86
10	40/4	10/3	4.95	2.80	3.87	0.89
11	40/4	10/2	4.60	2.87	3.74	1.04
12	40/4	10/1	4.60	3.00	3.80	0.89
13	40/3	10/3	4.45	2.30	3.37	1.05
14	40/3	10/2	4.40	2.48	3.44	0.84
15	40/3	10/1	4.18	2.87	3.52	0.93
16	40/2	10/3	3.20	1.98	2.59	0.80
17	40/2	10/2	3.35	1.98	2.66	0.88

*This standard deviation refers to listener mean ratings for the configurations. The listener mean is defined as the average rating given by an individual listener to a particular configuration. The standard deviations shown reflect the variations in the listener means for each configuration.

Config. No.	Mode 1 Frame/Cepstrum	Mode 2 Frame/Cepstrum	Male MCJ	Female MCJ	Overall MCJ	Std. Dev.*
18	40/2	10/1	3.15	2.30	2.73	0.87
19	20/4	20/3	6.48	5.18	5.82	0.86
20	20/4	20/2	6.55	4.93	5.74	1.07
21	20/4	20/1	6.20	4.55	5.37	1.12
22	20/3	20/3	6.52	2.95	4.74	1.21
23	20/3	20/2	5.80	3.40	4.60	0.96
24	20/3	20/1	5.18	3.43	4.30	1.31
25	20/2	20/3	4.65	2.27	3.46	1.07
26	20/2	20/2	4.37	2.45	3.41	0.99
27	20/2	20/1	4.65	2.57	3.61	1.02
28	20/4	10/3	7.07	5.10	6.09	0.90
29	20/4	10/2	6.35	5.00	5.68	0.85
30	20/4	10/1	6.85	5.45	6.15	0.72
31	20/3	10/3	6.07	3.73	4.90	0.75
32	20/3	10/2	6.02	3.62	4.82	1.02
33	20/3	10/1	6.07	3.68	4.87	0.89
34	20/2	10/3	4.70	2.52	3.61	0.76
35	20/2	10/2	4.73	2.80	3.76	1.18
36	20/2	10/1	4.30	2.02	3.16	1.09
37	40/4	40/4	3.87	3.30	3.59	1.32
38	40/3	40/3	4.20	1.50	2.85	1.15
39	40/2	40/2	2.35	0.95	1.65	0.72
40	40/1	40/1	1.23	1.07	1.15	0.75
41	10/4	10/4	3.25	5.23	4.24	0.92

Config. No.	Mode 1 Frame/Cepstrum	Mode 2 Frame/Cepstrum	Male MCJ	Female MCJ	Overall MCJ	Std. Dev.*
42	10/3	10/3	3.48	5.43	4.45	1.07
43	10/2	10/2	3.05	3.30	3.18	0.74
44	10/1	10/1	1.70	1.25	1.48	0.71
45	20/4	20/4	6.27	5.37	5.82	1.26
46	20/1	20/1	2.30	0.95	1.62	0.81
47	100,000 bps PCM		7.40	7.18	7.29	0.58

BIBLIOGRAPHY

1. J. L. Flanagan, Speech Analysis, Synthesis and Perception, Springer-Verlag, New York, 1972.

2. G. Fant, Acoustic Theory of Speech Production, Mouton and Company, The Hague, Netherlands, 1960.

3. H. Dudley, "Remaking Speech," Journal of the Acoustical Society of America, Vol. 11, pp. 169-177, 1939.

4. B. Gold and C. M. Rader, Digital Processing of Signals, McGraw-Hill Book Co., Inc., New York, 1969.

5. A. Papoulis, The Fourier Integral and Its Application, McGraw-Hill Book Co., Inc., New York, 1962.

6. J. Tierney, B. Gold, V. Sferrino, J. A. Dumanian, and E. Aho, "Channel Vocoder with Digital Pitch Extractor," Journal of the Acoustical Society of America, Vol. 36, pp. 1901-1905, 1964.

7. B. S. Atal and S. L. Hanauer, "Speech Analysis and Synthesis by Linear Prediction of the Speech Wave," Journal of the Acoustical Society of America, Vol. 50, pp. 637-655, 1971.

8. J. D. Markel and A. H. Gray, Jr., "A Linear Prediction Vocoder Simulation Based Upon the Autocorrelation Method," IEEE Transactions on Acoustics, Speech, and Signal Processing, Vol. ASSP-22, pp. 124-134, 1974.

9. A. V. Oppenheim and R. W. Schafer, "Homomorphic Analysis of Speech," IEEE Transactions on Audio and Electroacoustics, Vol. AU-16, pp. 221-226, 1968.

10. A. V. Oppenheim, "Speech Analysis-Synthesis System Based on Homomorphic Filtering," Journal of the Acoustical Society of America, Vol. 45, pp. 458-465, 1969.

11. A. M. Noll, "Cepstrum Pitch Determination," Journal of the Acoustical Society of America, Vol. 41, pp. 293-309, 1967.

12. J. Morton and A. Carpenter, "Experiments Relating to the Perception of Formants," Journal of the Acoustical Society of America, Vol. 35, pp. 475-480, 1963.

BIBLIOGRAPHY (Continued)

13. B. C. J. Moore, "Frequency Difference Limens for Short-Duration Tones," Journal of the Acoustical Society of America, Vol. 54, pp. 610-619, 1973.

14. D. A. Ronken, "Some Effects of Bandwidth-Duration Constraints on Frequency Discrimination," Journal of the Acoustical Society of America, Vol. 49, pp. 1232-1240, 1971.

15. K. N. Williams and D. R. Perrott, "Temporal Resolution of Tonal Pulses," Journal of the Acoustical Society of America, Vol. 51, pp. 644-647, 1972.

16. C. I. Malme, "Detectability of Small Irregularities in a Broadband Noise Spectrum," Quarterly Report, Massachusetts Institute of Technology, Research Laboratory of Electronics, January, 1959.

17. P. T. Brady, A. S. House, and K. N. Stevens, "Perception of Sounds Characterized by a Rapidly Changing Resonant Frequency," Journal of the Acoustical Society of America, Vol. 33, pp. 1357-1362, 1961.

18. B. O. Pyron and F. R. Williamson, "Study and Analysis of Speech Parameters and Bandwidth Compression Techniques," Final Report Contract DA-49-092-ARO-156, Army Research Office, Washington, D.C., June, 1967.

19. M. Halle, G. W. Hughes, and J. P. Radley, "Acoustic Properties of Stop Consonants," Journal of the Acoustical Society of America, Vol. 29, pp. 107-116, 1957.

20. A. M. Liberman, "Some Results of Research on Speech Perception," Journal of the Acoustical Society of America, Vol. 29, pp. 117-123, 1957.

21. H. Winitz, M. E. Scheib, and J. A. Reeds, "Identification of Stops and Vowels for the Burst Portion of /p,t,k/ Isolated from Conversational Speech," Journal of the Acoustical Society of America, Vol. 51, pp. 1309-1317, 1972.

22. D. J. Sharf and T. Hemeyer, "Identification of Place of Consonant Articulation from Vowel Formant Transitions," Journal of the Acoustical Society of America, Vol. 51, pp. 652-658, 1972.

23. K. N. Stevens and D. H. Klatt, "Role of Formant Transitions in the Voiced-Voiceless Distinction of Stops," Journal of the Acoustical Society of America, Vol. 55, pp. 653-659, 1974.

BIBLIOGRAPHY (Continued)

24. W. A. Ainsworth, "First Formant Transitions and the Perception of Synthetic Semi-vowels," Journal of the Acoustical Society of America, Vol. 44, pp. 689-694, 1968.

25. L. J. Raphael, "Preceding Vowel Duration as a Cue to the Perception of the Voicing Characteristics of Word-Final Consonants in American English," Journal of the Acoustical Society of America, Vol. 51, pp. 1297-1303, 1972.

26. P. Denes, "Effect of Duration on the Perception of Voicing," Journal of the Acoustical Society of America, Vol. 27, pp. 761-764, 1955.

27. W. A. Ainsworth, "Duration as a Cue in the Recognition of Synthetic Vowels," Journal of the Acoustical Society of America, Vol. 51, pp. 648-651, 1972.

28. G. W. Hughes and M. Halle, "Spectral Properties of Fricative Consonants," Journal of the Acoustical Society of America, Vol. 28, pp. 303-310, 1956.

29. A. W. F. Huggins, "Just Noticeable Differences for Segment Duration in Natural Speech," Journal of the Acoustical Society of America, Vol. 51, pp. 1270-1278, 1972.

30. A. W. F. Huggins, "On the Perception of Temporal Phenomena in Speech," Journal of the Acoustical Society of America, Vol. 51, pp. 1279-1290, 1972.

31. D. McNeill and B. Repp, "Internal Processes in Speech Perception," Journal of the Acoustical Society of America, Vol. 53, pp. 1320-1326, 1973.

32. P. Ladefoged and D. E. Broadbent, "Information Conveyed by Vowels," Journal of the Acoustical Society of America, Vol. 29, pp. 98-104, 1957.

33. J. C. Hammett, Jr., "An Adaptive Spectrum Analysis Vocoder," Ph.D. Thesis, School of Electrical Engineering, Georgia Institute of Technology, 1971.

34. B. Gold and C. M. Rader, "The Channel Vocoder," IEEE Transactions on Audio and Electroacoustics, Vol. AU-15, pp. 148-161, 1967.

35. A. C. Davies, "Correction of Zero-Order Hold Distortion in Digital Filters," IEEE Transactions on Audio and Electroacoustics, Vol. AU-19, pp. 289-292, 1971.

BIBLIOGRAPHY (Continued)

36. "IEEE Recommended Practice for Speech Quality Measurements," IEEE Transactions on Audio and Electroacoustics, Vol. AU-17, pp. 227-246, 1969.

37. J. B. O'Neal, Jr. and R. W. Stroh, "A Speech Encoder-Multiplexer Feasibility Study," Final Report, Air Force Office of Scientific Research, Contract F44620-70-C-0122, 1972.

38. B. J. McDermott, "Multidimensional Analysis of Circuit Quality Judgment," Journal of the Acoustical Society of America, Vol. 45, pp. 774-781, 1969.

39. E. H. Rothauser, G. E. Urbanek, and W. P. Pachl, "A Comparison of Preference Measurement Methods," Journal of the Acoustical Society of America, Vol. 49, pp. 1297-1308, 1971.

40. W. P. Pachl, G. E. Urbanek, and E. H. Rothauser, "Preference Evaluation of a Large Set of Vocoded Speech Signals," IEEE Transactions on Audio and Electroacoustics, Vol. AU-19, pp. 216-224, 1971.

41. J. C. Nunnally, Jr., Introduction to Psychological Measurement, McGraw-Hill Book Co., Inc., New York, 1970.

42. W. W. Hines and D. C. Montgomery, Probability and Statistics in Engineering and Management Science, The Ronald Press Co., New York, 1972.

43. J. L. Melsa, A. P. Sage, M. Srinath, and S. Jones, "Development of a Configuration Concept of a Speech Digitizer Based on Adaptive Estimation Techniques," Final Report, Defense Communication Agency, Contract DCA 100-72-C-0036, 1973.

44. T. P. Barnwell, III, private communication, Georgia Institute of Technology, 1974.

45. M. L. Uhrich, "Fast Fourier Transforms without Sorting," IEEE Transactions on Audio and Electroacoustics, Vol. AU-17, pp. 170-172, 1969.

46. J. I. Makhoul and J. J. Wolf, "Linear Prediction and the Spectral Analysis of Speech," Technical Report No. 2304, Bolt, Beranek, and Newman, Inc., Cambridge, Mass., 1972.

47. T. P. Barnwell, III, "An Algorithm for Segment Durations in a Reading Machine Context," Ph.D. Thesis, Massachusetts Institute of Technology, 1970.

48. T. P. Barnwell, III, J. E. Brown, III, A. M. Bush, and C. R. Patisaul, "Pitch Estimation for Speech Digitization," Final Report, RADC Post-Doctoral Program for the Defense Communication Agency, June, 1974.

VITA

Charles Richard Patisaul, the son of Charles Edward and Beulah Torrance Patisaul, was born October 23, 1946 in Milledgeville, Georgia. He married Nellie Carol Brannan of Lawrenceville, Georgia in August, 1969. They have one child, Charles Edward, II.

After graduation from Baldwin High School in Milledgeville, Mr. Patisaul entered Georgia Institute of Technology and received the B.E.E. degree with highest honor in 1969. He received the M.S.E.E. degree in 1970, at which time he entered the doctoral program. Upon completion of his degree requirements, he will assume the position of Lead Engineer with the Advanced Programs Department of Radiation, Inc., Melbourne, Florida.

He is a member of Phi Eta Sigma, Tau Beta Pi, Eta Kappa Nu, Phi Kappa Phi, and the Institute of Electrical and Electronic Engineers.